9th Annual Sequencing, Finishing, Analysis in the Future Meeting

Santa Fe, New Mexico
May 28-30, 2014

Link to agenda

Los Alamos
NATIONAL LABORATORY
EST. 1943

Contents

The 2014 "Sequencing, Finishing and Analysis in the Future" Organizing Committee:

* Chris Detter, Ph.D., Senior Science Advisor, DTRA
* Johar Ali, Ph.D., Reserach Director, International, alviarmani
* Patrick Chain, Bioinformatics/Metagenomics Team Leader, LANL
* Michael FitzGerald, Microbial Special Projects Manager, Broad Institute
* Bob Fulton, M.S., Director of Project Development & Management, WashU
* Darren Grafham, Lab Manager, Children's Hospital, Sheffield, UK
* Alla Lapidus, Ph.D., Director, Genomics, Algorithmic Biology Lab, SPbAU; Russia
* Donna Muzny, M.Sc., Director of Operations, BCM
* Nadia Fedorova, Genome Finishing and Analysis Team Leader, JCVI
* David Bruce, Project and Program Manager of Genomic Sciences, LANL
* Shannon Johnson, Ph.D., Project Manager, LANL
* Tracy Erkkila, M.S., Technical Project Manager, LANL
* Tina Graves, M.S., Reference Genomes, WashU

05/28/2014 - Wednesday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**Breakfast**	**x**	**American Breakfast Buffet**	**Sponsored by NEB**
8:30 - 8:45	Intro	**x**	Welcome Intro from Los Alamos National Laboratory	**TBD**
x	Session Chair	**x**	Session Chairs	Chair - Donna Muzny Chair - Bob Fulton
8:45 - 9:30	**Keynote**	FF0110	**Recent Advances in Cancer Genomics**	**Rick Wilson** **Sponsored by Sage Science**
9:30 - 9:50	Speaker 1	**FF0008**	Genomics on Google Cloud Platform: Store, Process, Explore and Collaborate	**Jonathan Bingham**
9:50 - 10:10	Speaker 2	**FF0086**	Rapid cloud-based data processing and analysis of >15,000 whole exomes in a collaborative setting promotes novel gene discovery	**Narayanan Veeraraghavan**
10:10 - 10:40am	**Break**	**x**	**Beverages and Snacks Provided**	**Sponsored by DNAnexus**
10:40 - 11:00	Speaker 3	**FF0083**	The Contiguity is Near --- PacBio	**Steve Turner**
11:00 - 11:20	Speaker 4	**FF0113**	illumina NGS update --- illumina	**Kelly Hoon**
11:20 - 11:40	Speaker 5	**FF0041**	The Ion PGM® Hi-QTM Sequencing Polymerase: Reducing Systematic Error, Increasing Accuracy, and Improving Read-length -- Thermo	**Anelia Kraltcheva**
11:40 - 12:20	**Panel Discussion**	**x**	**Next Generation Sequencing Technology Panel Discussion**	**Chair - Bob Fulton** **Chair - Patrick Chain**
12:20 - 1:45pm	**Lunch**	**x**	**Coronado Lunch Buffet**	**Sponsored by PacBio**
x	Session Chair	**x**	Session Chairs	Chair - Tina Graves Chair - Darren Grafham
1:45 - 2:00	Speaker 6	**FF0114**	Enabling Sequence-based Technologies for Clinical Diagnostics: FDA Division of Microbiology Devices Perspective	**Heike Sichtig**
2:00 - 2:15	Speaker 7	**FF0097**	FDA GenomeTrakr: building an international public heath lab network for foodborne pathogen tracking	**Ruth Timme**
2:15 - 2:30	Speaker 8	**FF0010**	The BCM-HGSC Clinical Exome: from concept to implementation	**Christian Buhay**
2:30 - 2:45	Speaker 9	**FF0099**	High Speed Variant Finding in Adenocarcinoma of the Lung using WGS	**Sterling Thomas**
2:45 - 3:00	Speaker 10	**FF0019**	Sequence analysis of plasmid diversity amongst hospital-associated carbapenem-resistant *Enterobactericeae*	**Sean Conlan**
3:00 - 3:15	Speaker 11	**FF0024**	Whole Genome Sequencing of Respiratory Viruses From Clinical Nasopharyngeal Swabs	**Darrell Dinwiddie**
3:15 - 3:30	Speaker 12	**FF0092**	Analyzing TB drug resistance	**Bette Korber**
3:30 - 4:00pm	**Break**	**x**	**Beverages and Snacks Provided**	**Sponsored by BioNano**
x	Session Chair	**x**	Session Chairs	Chair - Mike Fitzgerald Chair - Alla Lapidus
4:00 - 4:15	Speaker 13	**FF0018**	Universal Tail Amplicon sequencing for identification, characterization, classification & rare variant detection using biodefense and public health organisms	**Rebecca Colman**
4:15 - 4:30	Speaker 14	**FF0081**	Use of whole genome sequencing to determine the molecular mechanisms responsible for decreased susceptibility and resistance to azithromycin in *Neisseria gonorrhoeae*	**David Trees**
4:30 - 4:45	Speaker 15	**FF0007**	Transcriptional signatures in microbial diagnostics	**Roby Bhattacharyya**
4:45 - 5:00	Speaker 16	**FF0082**	Next generation sequencing (NGS) as an enhanced surveillance tool – the tale of *Salmonella enterica* serovar Heidelberg outbreaks associated with chicken consumption	**Eija Trees**
5:00 - 5:15	Speaker 17	**FF0031**	Towards Clinical utility in a next generation sequencing analytical pipeline	**Darren Grafham**
5:15 - 5:30	Speaker 18	**FF0108**	Genomics Capability Development and Cooperative Research	**Helen Cui**
5:30 - 5:45	Speaker 19	**FF0101**	HIV-1 Subtype Surveillance in Kenya: the puzzle of Emerging drug resistance and Implications on Continuing care	**Raphael Lihana**
5:45 - 6:00	Speaker 20	**FF0013**	Next Generation Sequencing Capability at NCDC – Lugar Center in Georgia	**Gvantsa Chanturia**
6:30 - 8:00pm	**Posters - Even #s** **Meet & Greet Party**	**EVEN #s**	**Poster Session with Meet & Greet Party (Sponsored by Roche)** **Food & Drinks**	**Sponsored by Roche** **6:30pm - 8:00pm**
8:00 - 9:30pm	**Posters - Odd #s** **Meet & Greet Party**	**ODD #s**	**Poster Session with Meet & Greet Party (Sponsored by Roche)** **Food & Drinks**	**Sponsored by Roche** **8:00pm - 9:30pm**
9:30 - bedtime	**on your own**	**x**	**Night on your own - enjoy**	**x**

05/29/2014 - Thursday

Time	Type	Abstract #	Title	Speaker
7:30 - 8:30am	Breakfast	x	La Fonda Breakfast Buffet	Sponsored by NEB
8:30 - 8:45	Intro	x	Welcome Intro from Los Alamos National Laboratory	TBD
x	Session Chair	x	Session Chairs	Chair - Johar Ali Chair - Donna Muzny
8:45 - 9:30	Keynote	FF0111	Dissecting the Missing Diagnostic Yield in exome sequencing	Deanna Church Sponsored by Personalis
9:30 - 10:00	Speaker 1	FF0057	A de novo Whole Genome Shotgun Assembler for Long Reads	Gene Myers
10:00 - 10:15	Speaker 2	FF0044	Speeding up NGS software development	D. Lavenier
10:15 - 10:30	Speaker 3	FF0038	New Frontiers of Genome Assembly with SPAdes 3.1	Anton Korobeynikov
10:30 - 11:00am	Break	x	Beverages and Snacks Provided	Sponsored by CLC Bio
11:00 - 11:15	Speaker 4	FF0065	Upgrading large genomes using Pacific Biosciences long reads and PBJelly software	Jeffrey Rogers
11:15 - 11:30	Speaker 5	FF0014	Assembly and Phasing of Polymorphic Heterozygous Diploid Genomes	Jason Chin
11:30 - 11:45	Speaker 6	FF0063	De novo Assembly of Medicago truncatula Genome Lines Using Illumina and Pacific Biosciences Sequencing Technologies	Thiruvarangan Ramaraj
11:45 - 12:00	Speaker 7	FF0062	Illumina sequencing with no artifacts	Zbyszek Otwinowski
12:00 - 12:15	Speaker 8	FF0034	Near perfect de novo assemblies of eukaryotic genomes using PacBio long read sequencing	James Gurtowski
12:15 - 1:55pm	Lunch	x	New Mexican Lunch Buffet	Sponsored by Promega
x	Session Chair	x	Session Chairs	Chair - Tina Graves Chair - Bob Fulton
1:55 - 2:10	Speaker 9	FF0069	Anchored Assembly: An Algorithm for Large Structural Variant Detection Using NGS Data	Niranjan Shekar
2:10 - 2:20	Speaker 10	FF0085	De Novo Assembly and Structural Analysis Using Extremely Long Single-Molecule Imaging	Han Cao
2:20 - 2:30	Speaker 11	FF0020	Human sequence assembly scaffolding using Irys genome maps	Heng Dai
2:30 - 2:45	Speaker 12	FF0071	Efficient de novo assembly of long NGS reads	Martin Simonsen
2:45 - 3:00	Speaker 13	FF0047	*de novo* mammalian assembly of one-library PCR-free 250-base Illumina reads	Iain MacCallum
3:00 - 3:15	Speaker 14	FF0032	Creating a Single Haplotype Human Genome Assembly	Tina Graves-Lindsay
3:15 - 3:45pm	Break	x	Beverages and Snacks Provided	Sponsored by BioNano
x	Session Chair	x	Session Chairs	Chair - Alla Lapidus Chair - Darren Grafham
3:45 - 4:00	Speaker 15	FF0033	Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)	Stephanie Guida
4:00 - 4:15	Speaker 16	FF0048	TE-Tracker: systematic identification of transposition events through whole-genome resequencing	Mohammed-Amin Madoui
4:15 - 4:30	Speaker 17	FF0035	Insights from analyzing *Clostridium botulinum* sequences	Karen Hill
4:30 - 4:45	Speaker 18	FF0037	AntibodyMining ToolBox: An Open Source Tool for the Rapid Analysis of Antibody Repertoires	Csaba Kiss
4:45 - 5:00	Speaker 19	FF0077	Genome evolution and GC patterns driven by recombination	Anitha Sundararajan
5:00 - 5:15	Speaker 20	FF0067	Comprehensive identification of structural variants in a robustly characterized personal human genome	William J Salerno
5:15 - 5:30	Speaker 21	FF0079	Methyl Sequencing, Dissecting the Subtleties of the Differentiated and Un-differentiated Genome	Masoud Toloue
6:00 - 8:00pm	Happy Hour(s)	x	Happy Hour at Cowgirl Cafe - Sponsored by illumina - Map Will be Provided	Sponsored by illumina
8:00 - bedtime	on your own	x	Dinner and Night on Your Own - Enjoy!!!	x

05/30/2014 - Friday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**Breakfast**	**x**	**Harvey House Breakfast Buffet**	**Sponsored by NEB**
8:30 - 8:45	Intro	**x**	Welcome Intro from Los Alamos National Laboratory	**TBD**
x	Session Chair	**x**	Session Chairs	Chair - Patrick Chain Chair - Alla Lapidus
8:45 - 9:30	**Keynote**	**FF0112**	**The fly biome: Dispersal of Human Pathogens by Airborne Mechanical Vectors**	**Stephan Schuster** **Sponsored by Advanced Analytic**
9:30 - 10:00	Speaker 1	**FF0006**	Multigenerational exposure to the Chernobyl environment in bank voles alters the mitochondrial genome	**Robert Baker**
10:00 - 10:15	Speaker 2	**FF0051**	Functional profiling of genomic fragments with Sequedex	**Ben McMahon**
10:15 - 10:30	Speaker 3	**FF0003**	Rare OTUs Reveal Oral Microbiome Seeding Relationship between Tsimane Mother-Infant Dyads Who Practice Premastication	**Joe Alcock**
10:30 - 11:00am	**Break**	**x**	**Beverages and Snacks Provided**	**Sponsored by Kapa Biosystems**
11:00 - 11:15	Speaker 4	**FF0053**	Optimization of Metagenomic Methods for TEDDY Microbiome Study	**Ginger Metcalf**
11:15 - 11:30	Speaker 5	**FF0002**	Assessing the sensitivity of viral metagenomics	**Nadim Ajami**
11:30 - 11:45	Speaker 6	**FF0061**	Viral Metagenome Pipeline	**Christian Olsen**
11:45 - 12:00	Speaker 7	**FF0091**	Recruiting human microbiome shotgun data to site-specific reference genomes	**Gary Xie**
12:00 - 12:15	Speaker 8	**FF0028**	Analysis of Mixtures Using Next Generation Sequencing (NGS) of Mitochondrial DNA: Forensic Applications	**Henry Erlich**
12:15 - 12:30	Speaker 9	**FF0075**	Selective Depletion of Abundant RNAs to Enable Transcriptome Analysis of Low Input and Highly Degraded RNA from FFPE Breast Cancer Samples	**Bradley Langhorst**
12:30 - 1:45pm	**Lunch**	**x**	**Santa Fe Deli Lunch Buffet**	**Sponsored by MRI**
x	Session Chair	**x**	Session Chairs	Chair - Mike Fitzgerald Chair - Darren Grafham
1:45 - 2:00	Speaker 10	**FF0029**	A Targeted Sequencing Approach to Enable Enhanced Sensitivity in Variant Detection	**Bob Fulton**
2:00 - 2:15	Speaker 11	**FF0058**	*De Novo* Mapping with Solid-State Detectors	**John Oliver**
2:15 - 2:30	Speaker 12	**FF0004**	Further improvements to Illumina library preparation from challenging samples	**Maryke Appel**
2:30 - 2:45	Speaker 13	**FF0095**	Technology advancements in large insert PacBio library construction for targeted sequencing	**Min Wang**
2:45 - 3:00	Speaker 14	**FF0089**	Tools of the trade: resolving repetitive and complex regions in genomes using next-generation sequencing technologies and manual genome finishing	**Aye Wollam**
3:00 - 3:15	**Closing Discussions**	**x**	**Closing Discussions for General Meeting - discuss next year's meeting………..Now go out and enjoy Santa Fe!**	**Chair - Chris Detter**

Introducing the evolution of NGS capture panels

xGen® Lockdown® Panels

Stocked Enrichment Panels for Targeted Next Generation Sequencing

xGen® Lockdown® Panels consist of individually synthesized and quality controlled Lockdown Probes that have been internally validated.

xGen® AML Cancer Panel v1.0

– 11,743 probes targeting 260 genes associated with the AML disease pathway.

xGen® Pan-Cancer Panel v1.0

– 7816 probes targeting 127 significantly mutated genes implicated across 12 tumor types.

xGen® Inherited Diseases Panel v1.0

– 116,355 probes targeting 4503 genes and 180 SNPs associated with inherited diseases.

IDT®
INTEGRATED DNA TECHNOLOGIES

05/28/2014 - Wednesday

Time	Type	Abstract #	Title	Speaker
7:30 - 8:30am	Breakfast	x	American Breakfast Buffet	Sponsored by NEB
8:30 - 8:45	Intro	x	Welcome Intro from Los Alamos National Laboratory	TBD
x	Session Chair	x	Session Chairs	Chair - Donna Muzny Chair - Bob Fulton
8:45 - 9:30	Keynote	FF0110	Recent Advances in Cancer Genomics	Rick Wilson Sponsored by Sage Science
9:30 - 9:50	Speaker 1	FF0008	Genomics on Google Cloud Platform: Store, Process, Explore and Collaborate	Jonathan Bingham
9:50 - 10:10	Speaker 2	FF0086	Rapid cloud-based data processing and analysis of >15,000 whole exomes in a collaborative setting promotes novel gene discovery	Narayanan Veeraraghavan
10:10 - 10:40am	Break	x	Beverages and Snacks Provided	Sponsored by DNAnexus
10:40 - 11:00	Speaker 3	FF0083	The Contiguity is Near --- PacBio	Steve Turner
11:00 - 11:20	Speaker 4	FF0113	illumina NGS update --- illumina	Kelly Hoon
11:20 - 11:40	Speaker 5	FF0041	The Ion PGM® Hi-QTM Sequencing Polymerase: Reducing Systematic Error, Increasing Accuracy, and Improving Read-length -- Thermo	Anelia Kraltcheva
11:40 - 12:20	Panel Discussion	x	Next Generation Sequencing Technology Panel Discussion	Chair - Bob Fulton Chair - Patrick Chain
12:20 - 1:45pm	Lunch	x	Coronado Lunch Buffet	Sponsored by PacBio
x	Session Chair	x	Session Chairs	Chair - Tina Graves Chair - Darren Grafham
1:45 - 2:00	Speaker 6	FF0114	Enabling Sequence-based Technologies for Clinical Diagnostics: FDA Division of Microbiology Devices Perspective	Heike Sichtig
2:00 - 2:15	Speaker 7	FF0097	FDA GenomeTrakr: building an international public heath lab network for foodborne pathogen tracking	Ruth Timme
2:15 - 2:30	Speaker 8	FF0010	The BCM-HGSC Clinical Exome: from concept to implementation	Christian Buhay
2:30 - 2:45	Speaker 9	FF0099	High Speed Variant Finding in Adenocarcinoma of the Lung using WGS	Sterling Thomas
2:45 - 3:00	Speaker 10	FF0019	Sequence analysis of plasmid diversity amongst hospital-associated carbapenem-resistant *Enterobactericeae*	Sean Conlan
3:00 - 3:15	Speaker 11	FF0024	Whole Genome Sequencing of Respiratory Viruses From Clinical Nasopharyngeal Swabs	Darrell Dinwiddie
3:15 - 3:30	Speaker 12	FF0092	Analyzing TB drug resistance	Bette Korber
3:30 - 4:00pm	Break	x	Beverages and Snacks Provided	Sponsored by BioNano
x	Session Chair	x	Session Chairs	Chair - Mike Fitzgerald Chair - Alla Lapidus
4:00 - 4:15	Speaker 13	FF0018	Universal Tail Amplicon sequencing for identification, characterization, classification & rare variant detection using biodefense and public health organisms	Rebecca Colman
4:15 - 4:30	Speaker 14	FF0081	Use of whole genome sequencing to determine the molecular mechanisms responsible for decreased susceptibility and resistance to azithromycin in *Neisseria gonorrhoeae*	David Trees
4:30 - 4:45	Speaker 15	FF0007	Transcriptional signatures in microbial diagnostics	Roby Bhattacharyya
4:45 - 5:00	Speaker 16	FF0082	Next generation sequencing (NGS) as an enhanced surveillance tool – the tale of *Salmonella enterica* serovar Heidelberg outbreaks associated with chicken consumption	Eija Trees
5:00 - 5:15	Speaker 17	FF0031	Towards Clinical utility in a next generation sequencing analytical pipeline	Darren Grafham
5:15 - 5:30	Speaker 18	FF0108	Genomics Capability Development and Cooperative Research	Helen Cui
5:30 - 5:45	Speaker 19	FF0101	HIV-1 Subtype Surveillance in Kenya: the puzzle of Emerging drug resistance and Implications on Continuing care	Raphael Lihana
5:45 - 6:00	Speaker 20	FF0013	Next Generation Sequencing Capability at NCDC – Lugar Center in Georgia	Gvantsa Chanturia
6:30 - 8:00pm	Posters - Even #s Meet & Greet Party	EVEN #s	Poster Session with Meet & Greet Party (Sponsored by Roche) Food & Drinks	Sponsored by Roche 6:30pm - 8:00pm
8:00 - 9:30pm	Posters - Odd #s Meet & Greet Party	ODD #s	Poster Session with Meet & Greet Party (Sponsored by Roche) Food & Drinks	Sponsored by Roche 8:00pm - 9:30pm
9:30 - bedtime	on your own	x	Night on your own - enjoy	x

NOTES

Speaker Presentations (May 28th)

Abstracts are in order of presentation according to Agenda

Keynote

FF0110

Recent Advances in Cancer Genomics

Rick Wilson

Washington University School of Medicine, The Genome Institute

NOTES

FF0008

Genomics on Google Cloud Platform: Store, Process, Explore and Collaborate

Jonathan Bingham

Product Manager, Genomics, Google Inc.

Generating genomics data is easier than ever before, but interpreting and analyzing it is still hard, and getting harder as the volume increases. Sequencing the whole genome of a single person produces more than 100 gigabytes of raw data, and a million genomes will add up to more than 100 petabytes. This abundance of new information carries great potential for research and human health -- and requires new standards, policies and technology. To contribute to the genomics community and help meet the data-intensive needs of the life sciences, Google recently introduced a simple web-based API to import, process, store, explore and collaborate with genomic data at “Google scale”. In conjunction with the Global Alliance for Genomics and Health as well as early collaborators in genomics and bioinformatics, we have begun to develop open source sample code and example analyses.

FF0086

Rapid cloud-based data processing and analysis of >15,000 whole exomes in a collaborative setting promotes novel gene discovery

Narayanan Veeraraghavan[1], Andrew Carroll[2], Shalini Jhangiani[1], Alexander Li[4], Tomasz Gambin[3], Zhuoyi Huang[1], Ginger Metcalf[1], Fuli Yu[1], Alanna Morrison[4], Donna Muzny[1], Richard Daly[2], James Lupski[3], Geoff Duyk[2], Richard Gibbs[1,3], Eric Boerwinkle[1,4]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA
[2]DNAnexus, Mountain View, CA, USA [3]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA [4]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA

We are now witnessing NGS-based efforts at consortium scales involving hundreds of collaborators and thousands of study subjects, to exhaustively characterize diseases and involved genes. There are three critical areas that serve as enablers for efficient genomic discovery and diagnostics: (a) large scale computational resources, (b) rapid and flexible analysis capabilities, and (c) efficient data delivery and consumption. To address these, we have entered into collaboration between DNAnexus, the Baylor-Hopkins Center for Mendelian Genomics (CMG), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, to create a Gene Discovery Commons for both Mendelian (CMG) and common chronic (CHARGE) diseases.

As a test pilot of the Commons, we analyzed 1000 cases from the CMG spanning 15 diseases, and 3500 whole genome and 11000 whole exome sequences of deeply phenotyped individuals from the CHARGE consortium. The CHARGE cohort samples provide a large comparison group for the CMG, thus increasing statistical power.

Rapid deployment of our Mercury NGS pipeline to Amazon cloud (AWS) was made possible through the DNAnexus platform. Some of the key features of the computational framework are: zero set-up, on-demand scale-up, version control, reproducibility, visualization, and compatibility with off-the-shelf third party tools, naturally enforced standards, compliance with CAP/CLIA/HIPAA, and ISO27001 data handling and security compliance. The ability to easily, quickly and securely share data between collaborators is yet another hallmark of this framework.

As an example of scientific utility, we sequenced a sample of 204 probands with heterotaxy, a syndrome involving malarrangement of internal organs within the chest and abdomen. Analysis of these data identified 38 genes including known heterotaxy genes and novel candidates.

These and other examples demonstrate the utility of the Gene Discovery Commons, and the effectiveness of a cloud base framework, for promoting collaboration, corroboration and clinical translation.

Keywords: cloud computing, data management, bioinformatics framework, gene discovery, collaborative science

FF0083

The Contiguity is Near

Steve Turner (sturner@pacificbiosciences.com), Pacific Biosciences

Single Molecule, Real-Time (SMRT®) Sequencing has advanced the state of the art in elucidating complex genomic regions of large genomes, which are characterized by highly repetitive, low-complexity regions and duplication events. This data can now be combined with transcriptome sequencing that has the ability to differentiate between the myriad isoforms that are difficult to resolve with short-read technologies. To date, the read lengths from SMRT sequencing have enabled highly contiguous genome assemblies for bacteria and model genomes. Now with improvements in assembly algorithms, and assembly polishing, Pacific Biosciences' long-read data have further led to high-quality assemblies that rival pre-second-generation clone-by-clone sequencing efforts. The recent release of new photo-protected dye chemistry (P5-C3) with N50 data >10,000 base pairs, has enabled whole-genome shotgun sequencing of larger genomes including arabidopsis, drosophila, spinach and the haploid human cell line (CHM1htert). These assemblies have resulted in many complete chromosome arms in a single contig. Furthermore, the contig N50s for these assemblies are typically megabases in length with one case representing a completed chromosome and have average base qualities of ~Q50. These data can be collected using a single library preparation method and do not require additional libraries such as jumping and mate pair library construction. Another area that benefits from long reads is full length transcript sequencing. Methods are now available for the construction of full length cDNA SMRTbell^{TM} libraries for single molecule sequencing. These have resulted in the discovery of novel isoform transcripts and have the additional value of long, intact reads which do not require an assembly approach for detection. These combined advances will aid in the understanding of large complex genomes and their functioning.

Keywords: Genome assembly, full length cDNA sequencing, long read sequencing

FF0113

Illumina NGS update

Kelly Hoon

Illumina, Inc

Illumina continues to drive improvements through new offerings in sample prep, instrumentation and analysis tools. The expanded portfolio aids in the continual effort to reduce run times and costs while expanding throughput offerings for diverse experimental needs. From amplicons to whole genome population studies and everything in between, Illumina has made great advances in 2014. This talk will provide an update on recent launches and those in the imminent future.

FF0041

The Ion PGM® Hi-Q^{TM} Sequencing Polymerase: Reducing Systematic Error, Increasing Accuracy, and Improving Read-length

Anelia Kraltcheva, Guobin Luo, Daniel Mazur, Mindy Landes, Sihong Chen, Kevin Heinemann, Theo Nikiforov, Joshua Shirley, Eileen Tozer, Peter Vander Horn

Thermo Fisher

High accuracy and sequencing uniformity are the hallmarks of a successful DNA sequencing platform. These qualities benefit greatly various scientific and medical applications. Here we show that the new Ion Torrent® Hi-Q^{TM} sequencing chemistry provides substantial improvements in overall accuracy, read-length, and systematic error relative to the previous sequencing chemistry. Using the Hi-Q^{TM} system we see a 50% decrease in overall error, mainly by reducing InDels, and up to a 90% decrease in systematic error. In addition, we show that with optimized emulsion PCR cycling conditions, we have rapidly increased coverage uniformity of the *Rhodobacter sphaeroides* genome from 88% to 95%. We also observed improvement in uniformity and representation of GC-rich amplicons from various human libraries.

By incorporating these improvements, the Ion PGM® system is now poised to enable a broader range of applications, such as enhanced de novo genome assemblies, Human Leukocyte Antigen (HLA) sequencing, bacterial identification, and meta-genomic analysis.

*For Research Use Only. Not for use in diagnostic procedures.

NOTES

Lunch

12:20 – 1:45pm

Sponsored by

PACIFIC
BIOSCIENCES™

NOTES

FF0114

Enabling Sequence-based Technologies for Clinical Diagnostics: FDA Division of Microbiology: Devices Perspective

Heike Sichtig
US FDA

The presentation will outline studies to evaluate the use of high throughput sequencing (HTS) devices as an aid in microbial diagnostics, and to gain a better understanding of the potential HTS clinical implementation strategies in microbial diagnostics. Focus will be on the possible approaches to validation studies and data for the evaluation of HTS systems for potential regulatory clearance/approval, and the use of sequence outputs from HTS devices to evaluate performance. Efforts towards generating an initial set of high quality, regulatory-grade microbial sequences through the FDA MicroDB project will also be discussed. Our vision is a robust, high quality microbial database that contains qualified regulatory-grade sequence data for use by developers and clinical end users. The information contained in the presentation concerning possible approaches for validation is not meant to convey FDA’s recommended approach.

Oral presentation Wed May 28th

FF0097

FDA GenomeTrakr: building an international public heath lab network for foodborne pathogen tracking

Ruth E. Timme, Peter Evans, Marc W. Allard, Errol Strain, Justin Payne, Christine Keys, Steven Musser

Food and Drug Administration, College Park MD USA.

The Center for Food Safety and Applied Nutrition at the FDA along with our partners, CDC and NCBI, are implementing an international NGS network of public health labs. Each lab collects and submits draft genomes of food pathogen to a reference database housed at NCBI. As new data arrives a public analysis pipeline reconstructs a phylogenetic tree for each pathogen database, providing the FDA actionable leads in outbreak investigations of foodborne pathogens. Currently the GenomeTrakr network comprises six state health labs (NY, VA, FL, AZ, WA, MN), nine FDA field labs, an FDA contracting lab, and CFSAN headquarters - each outfitted with one or more Illumina MiSeq instruments. This summer we are adding our first official international partner with Argentina. Herein, we report enhanced molecular epidemiological insights gained by comparative analysis of *Salmonella* and *Listeria* genomes, previously deemed indistinguishable by conventional subtyping methodologies. These results demonstrate an important investigative role for NGS tools within a regulatory environment while highlighting the novel additional insights provided to epidemiological investigations.

FF0010

The BCM-HGSC Clinical Exome: from concept to implementation

Christian Buhay[1], Qiaoyan Wang[1], Rashesh Sangvhi[1], Jianhong Hu[1], Yan Ding[1], Mark Wang[1], Dan Burgess[3], George Mayhew[3], Dawn Green[3], Yi Han[1], Huyen Dinh[1], Kimberly Walker[1], Harsha V. Doddapaneni[1], Yaping Yang[1], Eric Boerwinkle[1,2], Richard A. Gibbs[1] and Donna M. Muzny[1]

[1]Baylor College of Medicine. [2]Human Genetics Center, University of Texas Health Science Center at Houston, [3]Roche NimbleGen, Inc

For whole exome sequencing to become an integral part of routine clinical care, it must demonstrate high sensitivity and specificity, as well as aid in variant discovery and diagnoses across clinically relevant regions. We present the initial implementation of a medical whole exome capture reagent tailored to target genes implicated in over 3000 inherited disorders and various cancers.

Coverage data was examined from 34 whole exome sequencing (WES) samples in the Whole Genome Laboratory (WGL) at Baylor College of Medicine. Results from aggregate data reveal that utilization of the BCM-HGSC VCRome capture reagent fully covers (at ≥20X) over 75% of the medically relevant genes listed in COSMIC, HGMD, GeneTests, and ACMG Incidental Findings lists. Analysis also identified gene regions that were routinely below the 20X threshold. To address under-sequenced regions, the HGSC in collaboration with Roche/NimbleGen produced a small (220Kbp) targeted capture based reagent that is simply added to our existing VCRome exome solution. The ‘spike-in’ reagent targeted underperforming gene regions in GeneTests, ACMG, and 23 other clinical gene panels ranging from childhood cancers to neuromuscular disorders. Of the almost 2000 genes from the target list, the addition of top-up reagent improved coverage for ~550 genes. Ninety-seven percent of the genes are now fully covered (all bases ≥20X) from the previous baseline of 75%.

Ongoing collaborations with Nimblegen include expansion of the top-up reagent to include underperforming regions in OMIM and an expanded list of cancer genes. We surveyed over 3600 genes across 50 VCRome WES samples ranging in yield from 5Gbp to 7Gbp. Of the 3600 genes, 20K regions in 3100 genes were targeted for a 1Mbp design. Results from these ‘spike-in’ methods have broad clinical applications for insuring a high degree of sensitivity and specificity for variant calling across genes in various disorders of interest.

Keywords: Bioinformatics, Clinical, Genomics, Exome sequencing

FF0099

High Speed Variant Finding in Adenocarcinoma of the Lung using WGS

Sterling Thomas, Nathan Dellinger, Matt Feltz, Allan Bolipata, Danielle, Weaver, Tyler Barrus, Daniel Negron, Mitchell Holland

Noblis, Falls Church, Virginia
Nonprofit Science and Technology Organization

Advances in sequencing technology have increased the computational demands required for processing, identifying, and analyzing large and complex datasets. Because many genomic features such as Single Nucleotide Polymorphisms (SNPs) offer insight into carcinogenesis, origin of metastatic disease, and susceptibility to treatment, detecting these features rapidly within large sequence read sets is becoming increasingly valuable to the healthcare community.

To address these critical needs, Noblis' Center for Applied High Performance Computing (CAHPC), has developed a suite of high speed algorithms called BioVelocity that utilize the strengths of the CAHPC to perform reference-based multiple sequence alignment (MSA) and variant detection on human raw reads. This implementation uses next generation sequence reads as input and aligns them against a customized reference library, which can be specific – consisting of very similar tumors, or highly varied – containing tumors from multiple sites, stages and patient histories.

Adenocarcinoma of the Lung is a complex disease that afflicts both non-smokers and smokers. Since the cellular genesis of adenocarcinoma is different than small cell carcinoma, exome sequencing may not be sufficient to understand the genetic characteristics of the disease and it susceptibility to treatment. We are using over 400 whole genome sequencing (WGS) reads from paired tumor and normal samples that were submitted to The Cancer Genome Atlas (TCGA) to identify variants that may not be limited to the exome. To support this analysis we created a complete reference set supports access to all the genomes from the study, simultaneously, in active memory. This complete reference set allowed us to quickly identify variations that were unique to this population of tumors. This approach may be useful in creating genomic classification methodologies for complex diseases.

FF0019

Sequence analysis of plasmid diversity amongst hospital-associated carbapenem-resistant *Enterobactericeae*

Sean Conlan[1], Clayton Deming[1], Evan Snitkin[1], James C. Mullikin[2], Pam J. Thomas[2], Morgan Park[2], Jyoti Gupta[2], Shelise Y. Brooks[2], Brian Schmidt[2], Alice C. Young[2], Jim Thomas[2], Gerard G. Bouffard[2], Robert W. Blakesley[2], NISC Comparative Sequencing Program[2], Jonas Korlach[3], Tyson A. Clark[3], Khai Luong[3], Yi Song[3], Yu-Chih Tsai[3], Matthew Boitano[3], David Henderson[4], Karen M. Frank[4], Tara N. Palmore[4] and Julia A. Segre[2,1]

[1]National Human Genome Research Institute, Bethesda, MD, [2]National Institutes of Health Intramural Sequencing Center (NISC), Bethesda, MD, [3] Pacific Biosciences, Menlo Park, CA, [4]National Institutes of Health Clinical Center, Bethesda, MD.

We previously reported tracking the transmission of carbapenem-resistant *Klebsiella pneumoniae* amongst patients at the NIH Clinical Center, highlighting the importance of whole genome sequencing to understand the complexity of hospital transmission. Here we describe another layer of complexity to shape our understanding of outbreak surveillance and infection control; the repertoire of carbapenem-resistance encoding plasmids amongst the patient population and the hospital environment. The complex nature of multi-drug resistant bacterium, carrying many different plasmids which are themselves made up of mobile elements and gene cassettes, required improved tools for sequencing, finishing and annotation that exceed draft assembly pipelines. Sequencing and genomics technologies used in this study include PacBio RS II, Roche 454 FLX, Illumina MiSeq and the OpGen Argus. We analyzed patient and environmental hospital carbapenem-resistance Enterobactericeae, including *Klebsiella pneumoniae*, *Klebsiella oxytoca*, *Enterobacter cloacae* and *Citrobacter freundii.* Full genome sequencing revealed that these ten organisms carry carbapenem-resistance genes on eight different plasmids, challenging initial assumptions about horizontal gene transfer events. Indeed, the complex diversity of plasmids was only clarified with the PacBio multi-kilobase long reads assembled *de novo* into high-quality genomes, with the chromosome and each plasmid circularized. Data from other platforms were used to further qualify, characterize and make some corrections to the PacBio assemblies. The PacBio *de novo* assembly error rate was extremely low; typically 5 or less differences per genome, by inspecting these assemblies with high depth of coverage Illumina reads. These methods show it is now possible to better understand the transmission of bacterial strains and plasmids encoding antibiotic resistance using highly accurate and finished genome sequences of isolates from patients and the hospital environment.

FF0024

Whole Genome Sequencing of Respiratory Viruses From Clinical Nasopharyngeal Swabs

Dinwiddie, DL, Dehority, WN, Harrod, KS, Schroth, GP, Young, SA

University of New Mexico

Objectives/Specific Aims: Respiratory infections cause the greatest morbidity and mortality of all pediatric ailments accounting for ~20% of mortality for children younger than 5, one quarter of all hospitalizations, and between 33% and 59% of general practitioner consultations . Despite the enormous medical burden caused by respiratory viruses the specific genetic variation that influence transmission, virulence, and pathogenesis are poorly understand for most viruses. The objective of this study is identify genetic variation in clinical respiratory viruses that influence these critical processes.

Methods/ Study Population: We are developing a novel method to enrich viral genomes from clinical nasopharyngeal swabs, which will enable us to conduct deep whole genome sequencing on an Illumina MiSeq. We will conduct whole genome viral sequencing of respiratory syncytial virus, influenza A, human metapneuomivirus, and rhinovirus from residual nasopharyngeal swabs obtained during the peak 2013-2014 respiratory infection season in the state of New Mexico.

Results: We will use reference guided and *de novo* assembly approaches to characterize the complete viral genomes and identify genetic variation of individual clinical viral isolates. We anticipate that through bioinformatics analysis of the clinical viruses we will identify novel genetic variation that may effect transmission, virulence and pathogenesis of the virus.

Discussion/Significance of Impact: Our results will provide crucial insight into the genomic diversity of clinical viral strains, pathophysiology of infection, and may provide understanding into causes of resistance to antiviral therapies.

FF0092

Analyzing TB drug resistance

Karina Yusim, Shihai Feng, Taeksun Song, Clif Barry, Bette Korber

Los Alamos National Laboratory

Drug resistant bacteria are a growing challenge of 21st century. Extensively drug resistant (XDR) TB is on the rise and is present in 92 countries, including US. XDR TB (resistant to most potent drugs and difficult to diagnose) is contagious, airborn and lethal. Conventional culture-based phenotypic drug susceptibility testing is slow, and results are inaccurate for some drugs. Available genetic tests, line probe assays and the molecular beacon GeneXpert platform, are very rapid and sensitive, but can only sample a limited number of alleles. Despite these difficulties, drug susceptibility testing is important. If an the infecting strain is resistant to the treatment, the patient is left vulnerable, and treatment with a multidrug therapeutic cocktail that is only partially effective can result in acquisition of resistance to the other drugs in the cocktail, failed therapy and spread of resistant forms. I will discuss how the whole genome sequencing and analysis can help close this technology gap. I will also present a collaborative study between the groups in S. Korea, NIAID and Los Alamos, where we identified a striking example of compensatory mutations through whole genome analysis; these mutations restore fitness, and are associated with transmission of rifampicin-resistant TB. The computational tools we developed and the extensive knowledge base of *Mycobacterium Tuberculosis* drug resistance mutational pathways can be leveraged and can be used as a model system of bacterial drug resistance.

FF0018

Universal Tail Amplicon sequencing for identification, characterization, classification and rare variant detection using biodefense and public health organisms

Colman RE[1], Schupp JA[1], Sahl J[1], Hicks ND[1], Smith DE[1], Valafar F[3], Rodwell TC[4], Catanzaro A[4], Wagner DM[2], Keim PS[1,2], Engelthaler DM[1]

1. Translational Genomics Research Institute, Pathogen Genomics Division, Flagstaff, AZ 2. Ctr. for Microbial Genetics & Genomics, Northern Arizona Univ., Flagstaff, AZ 3. San Diego State Univ., San Diego, CA 4. Univ. of California San Diego, San Diego, CA

Applying NGS directly to forensics or clinical samples currently has limitations due to the complexity of biological samples. Using a novel amplicon NGS methodology allows for near-real time identification, characterization, classification and, with certain applications, rare variant detection. Employing a universal amplicon indexing system, we prepared flow-cell ready amplicons, from multiplexed PCRs. This amplicon approach allows for fast screening from complex samples across a wide array of targets. We present two applications of this universal tail amplicon NGS methodology: A) a bioforensics application with *Burkholderia pseudomallei,* and B) a clinical rare variant application in *Mycobacterium tuberculosis.* We developed *B. pseudomallei* amplicons based on MLST targets, and targets identified in house for the identification, characterization, and classification of *B. pseudomallei*. The use of targeted amplicon sequencing allows for definitive characterization on low quality or quantity samples, which is important in a forensic setting. Rapidly characterizing an important biodefense organism allows for comprehensive information collected for a sample to help direct investigations or clinical decisions. We also developed amplicons based on drug resistance (DR) conferring SNPs for rare variant detection in *M. tuberculosis*. The use of universal tail amplicon methodology and "single molecule-overlapping reads" (SMOR) analysis for determination of actual mutation ratios in target loci leads to an increase in mixed population detection sensitivity and lower probability of erroneous base calls. Undetected extremely low level sub-populations of DR in the *M. tuberculosis* populations (i.e., heteroeresistance) at patient treatment initiation may play a role in the development of DR-TB. Current minor DR detection levels are limited to above 1%, at best. We have developed a method of detecting resistant *M. tuberculosis* sub-populations consisting of 0.1% of the total *M. tuberculosis* population in almost real time. These approaches significantly improve on our ability to analyze previously undetectable genomic features of forensic and clinical samples.

Keywords: amplicon sequencing, *Burkholderia pseduomallei*, *Mycobacterium tuberculosis*, Miseq, rare variant

FF0081

Use of whole genome sequencing to determine the molecular mechanisms responsible for decreased susceptibility and resistance to azithromycin in *Neisseria gonorrhoeae*

David Trees[1], Serena Carroll[1], John Papp[1], Kevin Karem[1], Yonatan Grad[2,3], Steve Johnson[1]

[1] Division of STD Prevention, NCHHSTP, Centers for Disease Control and [2] Center for Communicable Disease Dynamics, Department of Epidemiology, Harvard School of Public Health. [3]Division of Infectious Diseases, Brigham and Women's Hospital, Harvard Medical School

Objective: *Neisseria gonorrhoeae* has ability to rapidly develop resistance to each antimicrobial class used for therapy, including the currently recommended co-treatment antimicrobial azithromycin. Emerging resistance to azithromycin will complicate clinical and public health management of gonorrhea. Molecular approaches, such as whole genome sequencing to identify mechanisms of resistance and their prevalence, can be useful for the development of point-of-care molecular assays to detect resistance or treatment failure. We have sequenced *N. gonorrhoeae* isolates with elevated azithromycin minimum inhibitory concentrations (MICs) to identify mutations that confer resistance or increased MICs.

Methods: Strains with elevated MICs to azithromycin were obtained through the Gonococcal Isolate Surveillance Project coordinated by CDC. DNA was isolated and whole genome sequencing was performed on Illumina and Pacific Biosciences platforms. WGS data was analyzed using CLCBio software. Sanger sequencing confirmed the presence and number of 23S rRNA allelic variants.

Results: Analysis of WGS demonstrated a wide variety of mutations associated with increased MICs to azithromycin. Mutations in the multiple transferable resistance repressor (mtrR) loci, including an insertion of meningococcal DNA, were found in numerous locations within the promoter and gene regions. Additionally, a 23S rRNA variant, C2599T, known to be associated with macrolide resistance, occurred in 2, 3, or 4 of the four 23S rRNA alleles present in a majority of the isolates with elevated MICs to azithromycin, and was not found in any susceptible isolate.

Summary: Analysis of the WGS data revealed 2 predominant mechanisms of resistance to azithromycin in *N. gonorrhoeae* in isolates in the USA: 1) mutations in either the promoter region or structural gene of mtrR or 2) mutations in at least 2 of the 23S rRNA alleles located on the gonococcal chromosome. This information may prove useful in the development of rapid, culture-independent molecular based tests for the detection of treatment resistant gonococci.

Keywords: *Neisseria gonorrhoeae*, antimicrobial resistance, azithromycin, whole genome sequencing

FF0007

Transcriptional signatures in microbial diagnostics

Roby P. Bhattacharyya[1,2] (rbhatt@broadinstitute.org), Jonathan Livny[1], Robert Rudy[1], Milesh Patel[1], Deborah T. Hung[1,3, 4, 5]

[1]Infectious Disease Initiative, Broad Institute of MIT and Harvard, Cambridge, MA, USA; [2]Department of Medicine, Division of Infectious Disease, and [3]Department of Molecular Biology, Massachusetts General Hospital, Boston, MA, USA; [4]Department of Medicine, Division of Infectious Disease and Critical Care, Brigham and Women's Hospital, Boston, MA, USA; [5]Department of Microbiology and Immunology, Harvard Medical School, Boston, MA, USA

Rapid antibiotic administration remains our most effective weapon against bacterial pathogens, but recent emergence of antibiotic resistance in hospitals and the community poses new challenges. Current methods of culture-based susceptibility determination are too slow for early evidence-based selection of effective antibiotics in seriously ill patients, or for routine use in outpatieint clinical settings. RNA detection has the potential to identify antibiotic susceptibility more rapidly than current culture-based diagnostic methods. RNA expression patterns change within 10 minutes in bacterial pathogens upon exposure to antibiotics. Further, our lab has shown that susceptible bacteria enact different transcriptional programs than resistant ones upon antibiotic exposure, independent of the mechanism of resistance. This suggests the possibility of an RNA-based phenotypic assay for antibiotic resistance that does not require prior knowledge of the genetic basis for this resistance.

We report a transcriptomic analysis of numerous key drug-resistant pathogens (the *ESKAPE* organisms and *M. tuberculosis*) against a variety of antibiotics using high-throughput RNA-Seq. These studies reveal transcriptional signatures of antibiotic susceptibility and allow us to identify the transcripts whose expression levels most clearly distinguish susceptible from resistant organisms within minutes of drug exposure. We validate these signatures using a rapid commercial RNA hybridization assay (Nanostring), which allows us to discriminate between susceptible and resistant clinical isolates within hours instead of days, as current methods require. This RNA-based technique offers a novel approach to diagnosing antibiotic resistance that bridges genotype and phenotype, with the potential for unprecedented speed that would transform clinical infectious disease diagnostics.

Keywords: RNA-Seq, antibiotic resistance, molecular diagnostics

FF0082

Next generation sequencing (NGS) as an enhanced surveillance tool – the tale of *Salmonella enterica* serovar Heidelberg outbreaks associated with chicken consumption

Eija Trees, Ashley Sabol, Dana Castillo, Beth Tolar, Efrain M. Ribot, and Heather Carleton

Centers for Disease Control and Prevention, Atlanta, GA

Salmonella enterica serovar Heidelberg causes over 100,000 illnesses annually in the USA, a high proportion of which are attributed to poultry consumption. PFGE -based surveillance of serovar Heidelberg is somewhat limited because 42% of isolates in the PulseNet national database exhibit an indistinguishable restriction pattern (pattern 22) highlighting the need for a more discriminatory subtyping method. During the summer of 2013 the national prevalence of pattern 22 rose above background. Three serovar Heidelberg outbreaks were identified during this period, two of which (outbreaks 1 and 2) were associated with social functions where meals containing chicken were served, and one (outbreak 3) was a long-term multi-state event involving seven PFGE patterns, including pattern 22, and associated with chicken from multiple production facilities of a single manufacturer. It was unclear whether there was any connection between the chicken involved in these events due inconclusive trace-back investigations. In this study genomic sequencing was used to determine if additional strain discrimination would help resolve potential epidemiological connections. Forty-six isolates were sequenced with 2x150bp chemistry on the Illumina Miseq. Assemblies and hqSNP calls (20x coverage, 95% frequency) were done using the CLCBio Genomics Workbench. Outbreaks 1 and 2 appeared not related, with an average of 110 hqSNPs separating them. Isolates within the outbreaks clustered tightly together with less than 15 hqSNPs and bootstrap values of 100. Isolates belonging to outbreak 3 showed more variation even between isolates within the same PFGE pattern. The two pattern 22 isolates from outbreak 3 differed from each other by 56 hqSNPs and from outbreaks 1 and 2 by an average of 60 and 100 SNPs, respectively. These results suggest that the three outbreaks were most likely not caused by a common source, and that the higher variation within outbreak 3 likely reflects the multi-source contamination within the production facilities.

Keywords: *Salmonella,* Heidelberg, chicken, sequencing, hqSNP

FF0031

Towards Clinical utility in a next generation sequencing analytical pipeline

Darren Grafham, Kevin Blighe

Sheffield Diagnostic Genetics Service

Sheffield Diagnostic Genetics Service is the most integrated NHS diagnostic genetics department in the UK offering over 90 diagnostic tests covering both cytogenetic and molecular genetic disciplines to the UK and internationally. As next generation sequencing (NGS) technology has become a regular fixture in research, it is now time for it to take the stage in clinical environments in order to be regularly used in diagnostics and screening. However, there is still a shadow of doubt about the widespread use of NGS technology as a diagnostic tool. Before implementation, there needs to be a consensus on which analytical pipeline to use, alongside follow-up confirmation of variants with the current gold standard in diagnostics :Sanger sequencing.

Here, we present an NGS analytical pipeline that has complete agreement on 341 variants with Sanger sequencing and that is being used in a clinical diagnostic laboratory for regular screening of inherited, pathogenic variants. The novel approach in this analytical pipeline, which involves randomly selecting subsets of reads and later merging variants from each, allows for false-negatives to be eliminated and ensures complete pick-up of variants. Moreover, modelling reduced depth of coverage reveals that 30X is the point at which false-positives are equally eliminated with >99.9% confidence.

FF0108

Genomics Capability Development and Cooperative Research with Global Engagement

Helen Cui (hhcui@lanl.gov), Tracy Erkkila, Patrick Chain

Los Alamos National Laboratory, Bioscience Division, NM 87544

Genomics science and technologies are transforming life sciences globally, and becoming a highly desirable international collaboration area. Los Alamos National Laboratory is leveraging our own long term research and development experience and expertise to assist multiple countries and regions in advancing genomics capabilities, focusing on genomics scientific foundation, next generation sequencing technology, and analytics in pathogen detection and characterization and biosurveillance applications. Our current partner countries and regions include Republic of Georgia, Jordan, Kenya, Yemen, Gabon, Uganda, and South East Asia; collaborations with other countries and regions are being developed.

With these collaborations, we are assisting the host nations to develop the capabilities that are urgently needed to address pressing challenges in infectious disease spreads, implementing safe laboratory practices, and developing infectious disease detection and characterization techniques that can be maintained and further developed by the host countries. LANL has developed bioinformatics pipelines that enable the partner countries to process and analyze next generation sequencing data output. Such collaboration efforts not only benefit the host countries and region with the state-of-the-art life science and technologies, but also build a trusted international community with shared passion in addressing global emerging infectious challenges and shared resources, which are essential for approaching effective global health objective and meeting International Health Regulation requirements.

FF0101

HIV-1 Subtype Surveillance in Kenya: the puzzle of Emerging drug resistance and Implications on Continuing care

Raphael W. Lihana, Michael Kiptoo, Elijah M. Songok

Kenya Medical Research Institute, Nairobi, Kenya

Background HIV-1 is characterized by genetic diversity such that specific viral subtypes are predominant in specific geographical areas worldwide. The genetic variation in HIV-1 genes is responsible for rapid development of resistance to current drugs. Though antiretroviral therapy (ART) has played a major role in reducing the impact of HIV/AIDS, in resource-limited settings, poor infrastructure, treatment interruption, bad adherence and drug stock-outs have led to emergence of drug resistance mutations in infected populations. To counter this, it has become vital to determine genetic subtypes, baseline as well as acquired mutations associated with drug resistance among the circulating strains. In Kenya, the distribution of HIV-1 subtypes has shown continuous evolution with different geographic regions as well as populations harboring different subtypes. As a continuing HIV-1 surveillance in Kenya, the prevalence of HIV-1 subtypes and drug resistance mutations were determined with the aim of localizing individual subtypes and mutations to different populations across the country.

Methodology Patient samples were collected from three sites with distinct cultural backgrounds in Kenya between 2006 and 2011. These included Kitale, Kapsabet and Nandi hills in Rift valley, Kenyatta National hospital in Nairobi and Coast provincial general hospital in Mombasa at the Kenyan coast. Demographic data was collected using a self-administered questionnaire. Whereas some patients were on ART, majority were naive. From each individual, 5ml of blood was obtained. CD4^{+}T-cell count and plasma HIV-1 RNA load (VL) were determined according to kit manufacturers' instructions. HIV-1 RNA was extracted from 140µl of plasma using Qiagen kit. An initial One-Step RT-PCR was performed targeting a portion of HIV-1 *pol* gene using the One-Step RT PCR Kit followed by a nested PCR amplification using in-house group M primers. Amplicons were sequenced employing the BigDye chemistry (Applied Biosytems, Foster City, CA, USA). Generated sequences were aligned using ClustalW and phylogenetic trees inferred using FigTree. Nucleotide sequences were translated into the corresponding amino acids and analyzed for previously reported drug resistance-associated mutations using the IAS algorithm and the Stanford University HIVdb analysis program.

Results A total of 312 patient samples were analyzed [59 from coast region (drug experienced), 188 from Rift valley region (drug naïve) and 65 from central region (drug experienced)]. Of these, 262 were females while 50 were males. Prevalent HIV-1 subtypes included: A1 (210, 67.3%), A2 (2, 0.6%), C (22, 7.1%), D (48, 15.4%), G (4, 1.3%), AC (3, 1%), AD (20, 6.4%) and CD (3, 1%). The prevalence of drug resistance mutations was 3.2%, 26% and 27% in Rift valley, Central and Coast regions, respectively.

Conclusion Majority of HIV-1 subtypes in Kenya were A1, C and D. The proportion of subtype D was highest in Rift valley and lowest in the coast region. Possible HIV-1 recombinants accounted for 9% (28/312) of the samples. Major mutations were associated with current regimens in Kenya. Future ART in Kenya would require proper laboratory infrastructure to monitor evolving subtypes as well as guide treatment using available regimens.

FF0013

Next Generation Sequencing Capability at NCDC – Lugar Center in Georgia

Gvantsa Chanturia

National Center for Disease Control and Public Health of Georgia, Lugar Center for Public Health Research

The National Center for Disease Control and Public Health of Georgia(NCDC) – the former anti-plague station(APS) is the main institution in the country of Georgia that carries out active surveillance on endemic viral and bacterial vector-borne zoonotic diseases, as well as respiratory and enteric outbreaks since 1937. Over the years, clinical and environmental samples were investigated for the presence of various pathogens and now serve as a large National endemic pathogen collection.

In the scope of our scientific projects, funded by different donors, and in particular by the DTRA, significant improvements in local research technologies have been made over the last several decades. In addition to established classical bacteriological and serological methods, one by one the following methods have been added: molecular detection, genotyping and Sanger sequencing. Next Generation Sequencing (NGS) Illumina MiSeq platform and BSL-3 facility have became the part of NCDC at 2013, after incorporation of R.G.Lugar Center for Public Health Research.

Los Alamos National Laboratory is now assisting in development of NGS and analytic capabilities. Several training sessions have occurred both on-site and remotely for DNA library preparation, sequencing, data analysis and other bioinformatics. Selected strains of *B. anthracis* and *F. tularensis* were already sequenced. Ten *Brucella* diagnostic phages were sequenced and analyzed in collaboration with the Eliava Institute of Bacteriophages. Ilia State University has also used the NGS facility for phylogenetic characterization of endemic species of rock lizards.

The laboratory is now actively moving forward in developing new scientific ideas. The new projects for characterization of the NCDC strain repository using NGS and investigating both clinical & environmental samples using state-of-the-art metagenomic approaches are being planned in the near future. Collaborations with other institutions and universities within Georgia and also neighboring countries will be one of the main goals of the facility in the future.

Keywords: Georgia, NGS facility, DTRA

Notes

Meet and Greet Party

630pm – 930pm, May 28th

Sponsored by Roche Diagnostics

Enjoy!!!

Poster Presentations (May 28^{th})
Even #'s 6:30-8:00pm, Odd #'s 8:00-9:30pm

FF0005

qPCR-based quantification and QC assays provide actionable data for NGS library construction from FFPE samples of variable quality

Maryke Appel[1], Mike Becker[2], Michael Berger[3], Corinne Camalier[4], Luis Carvajal-Carmona[5], Lisa Cook[2], Matt Cordes[2], Robert Daber[6], Biswajit Das[4], Tracie DeLuca[2], Ryan Demeter[2], Catrina Fronick[2], Bob Fulton[2], Lucinda Fulton[2], Kety Huberman[7], Vincent Magrini[2], Sean McGrath[2], Paul McGregor[4], Deborah Poruban[8], Kelsi Rotter[2], Ruta Sahasrabudhe[5], Sasinya Scott[3], Martina Veigl[8] and John Foskett[1]

[1]KAPA Biosystems, Inc., Wilmington, MA [2]The Genome Institute, Washington University in St. Louis, St. Louis, MO [3]Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, NY [4]National Cancer Institute at Frederick, Frederick, MD [5]Genome Center and Department of Biochemistry and Molecular Medicine, University of California, Davis, CA [6]Clinical Genomics, Center for Personalized Diagnostics, University of Pennsylvania, Philadelphia, PA [7]Geoffrey Beene Translational Oncology Core, Memorial Sloan-Kettering Cancer Center, New York, NY [8]Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH

Formalin-fixed paraffin embedded (FFPE) tissue is an important source of DNA for cancer genomics studies and clinical diagnostics. DNA extracted from FFPE samples is, however, typically limited to nanogram quantities, and riddled with molecular damage that further lowers usable input. This impacts the ability to construct high-quality libraries for next-generation sequencing. One of the major challenges of high-throughput sequencing is the ability to process low-input samples of variable quality with predictable success rates in standard sample preparation pipelines.

Electrophoretic profiles and/or spectrophotometric quantification data for FFPE DNA are typically inadequate to predict library yields and quality. Kapa Biosystems produces qPCR-based assays that allow for the quantification and quality assessment of input DNA, and accurate quantification of adapter-ligated libraries after ligation and (pre-capture) amplification. These assays provide actionable data for FFPE library construction.

In the first part of this study, the KAPA Library Quantification Kit was used to assess two key library construction metrics for libraries constructed from FFPE DNA or high-quality control DNA. Conversion rate (% input DNA converted to adapter-ligated library), rather than amplification efficiency, proved to be the major bottleneck in FFPE library construction, with FFPE samples yielding 60 – 85% less adapter-ligated library than high-quality DNA. This limits library diversity, and leads to high duplication rates.

In the second part of the study, quality scores (Q129/41-ratios) for FFPE DNA, generated using the KAPA Human gDNA Quantification and QC Kit, were correlated with key sequencing metrics (coverage and duplication rates), for two sets of FFPE capture libraries sequenced on the Illumina platform. The data indicated that samples with a Q129/41-ratio >0.4 can be processed in standard target capture pipelines with a high chance of success. Higher input, or lower output requirements, rather than more amplification cycles, are recommended for samples with a Q129/41-ratio <0.4.

FF0009

Unsupervised Phylogeny with Automatic Correction for Horizontal Gene Transfer

Raquel Bromberg, Zbyszek Otwinowski

University of Texas, Southwestern Medical Center, Dallas, TX

Advances in sequencing are generating a rapidly increasing number of genomes. Alignment-based methods may not necessarily scale with the growth of information; moreover, they need a list of orthologues, separated from paralogues generated by horizontal transfer and gene duplication followed by selective loss. We have developed a method free from dependence on such considerations, which produces phylogenies constructed from whole proteomes and which is more robust than past efforts in this category. It is a unique alignment-free method that can identify horizontal gene transfer (HGT) and correct for its presence. The SlopeTree method automatically and efficiently calculates genomic evolutionary distances, with robust phylogenic inference even when using a fast, Neighbour-Joining method. SlopeTree calculates evolutionary distances between organisms using the statistics of exact kmer matches between proteomes, and hierarchically corrects for the effects of horizontal gene transfer. Comparison of SlopeTree results with the NCBI taxonomy and trees produced by conserved protein concatenation validates the concept of bacterial species and phyla as having a core proteome evolving by descent.

FF0011

The dynamics of Influenza Isolates in Uganda, their implications and way forward

Timothy Byaruhanga, J T Kayiwa, I. Nabukenya, R. Chiza, B. Namagambo, B. Bakamuntumaho, J.J Lutwama

Uganda Virus Research Institute, Entebbe/UG

Background
The isolation of Influenza viruses is important in detection of circulating strains in surveillance programs and development of vaccines. The fact that Influenza B viruses are rare in animals and there is no known animal reservoir and thus poses negligible pandemic threat has caused a decline in information concerning Influenza B in comparison to Influenza A viruses. We analyzed the changes that have occurred in Influenza A and B virus isolates in Uganda from November 2011 to May 2013 to determine extent of changes that have occurred in the viruses.

Method
Clinical ILI and SARI samples positive for Influenza viruses by Polymerase Chain Reaction (PCR) were inoculated and propagated on Madin-Derby Canine Kidney (MDCK) cell line. Hemmaglutination and Hemmaglutination inhibition (HA/HAI) test was carried out using guinea pig and turkey erythrocytes. The isolates were tested for drug resistance, sequenced and phylogenetically characterized.

Results
Out of 3455 samples collected, 439 samples were positive for Influenza A and B, with 324 Influenza A and 115 Influenza B. Only 41 (12.6%) Influenza A isolates were recovered with 33 isolates (80.5%) being recovered at passage two. Of the 115 Influenza B, 57 (49.6%) isolates were recovered, 36 isolates (63.2%) being recovered at passage one. Influenza B viruses were classified as group 1 B/Brisbane/60/2008-like with amino acids changes at positions N75K, N165K and S172P of the HA gene. AH3N2 HA gene was classified in subgroup 3B and 3C, with subgroup 3B having amino acid changes at N145S and subgroup 3C at S45N (glycosylation) and T48I. Influenza A viruses showed decreased HA titre at 16HA units but had constant HAI titre at 1280, B viruses showed higher HA titre at 128HA and a decreased HAI titre at 640. Influenza B and A/H3N2 viruses were sensitive to the antiviral drugs Oseltamivir and Zanamivir.

Conclusion
Influenza B viruses are easier to isolate using MDCK cells than Influenza A viruses. There are no significant changes in current circulating strains to cause a pandemic threat.

Key Words
Influenza, Isolation, Sequencing, Drug Resistance, Pandemic threat

FF0012

Comparison between benchtop next generation sequencers Illumina MiSeq and on Torrent PGM for genome assembly, SNP calling, and wgMLST in *Listeria monocytogenes*

Heather A. Carleton-Romer, Patti Lafon, Ashley Sabol, Katie Roache, Hannes Pouseele[1], Cheryl Tarr, Peter Gerner-Smidt, Efrain M. Ribot, Eija Trees

Centers for Disease Control and Prevention, Atlanta, GA; [1]Applied Maths, Sint-Martens-Latem, Belgium

Introduction: The use of next generation sequencers (NGS) is increasing in public health laboratories, but it is unclear if data generated by differing platforms can be used interchangeably in a single surveillance system such as PulseNet. In this study, sequences generated on the Illumina MiSeq and Ion Torrent PGM platforms from the same *Listeria monocytogenes* strains were compared to determine if there were differences in assembly metrics, or in high quality (hq)SNPs and whole genome multi-locus sequence typing (wgMLST) allele calls.

Methods: Sequencing was performed for 22 isolates using a 2x150bp sequencing chemistry on the MiSeq and 200bp chemistry on the PGM. Sequence data from the PGM was assembled using MIRA and for the MiSeq reads CLCBio Genomics Workbench was used. All SNPs were called at ≥10x coverage and ≥75% frequency using CLCBio. SNP calls in homopolymers were filtered out of the PGM data. SNP calls were compared using the SNP Extraction Tool (SET), and wgMLST data was generated using BioNumerics 7.5.

*Results***:** Comparing platforms, the average coverage for the genome sequences generated on the PGM was 47x (range: 21x - 73x) and for the MiSeq 128x (58x - 266x). Overall, genome sequences generated on the MiSeq assembled into fewer contigs (an average of 22 versus 28 on the PGM) and the N50 was higher (391,927 versus 306,604 for the PGM). For the hqSNP calls, sequences for the same isolate differed by 0-2 hqSNPs when comparing calls by the platform. For wgMLST, there were on average 16 more loci identified from the Miseq sequence data and 0-2 discrepancies when comparing assembly-free allele calls for the same isolate between platforms.

Conclusion: This preliminary analysis suggests that data from the MiSeq and PGM are compatible for use in a single surveillance system. Further evaluation is in process for *Salmonella* and *E. coli*.

Keywords: MiSeq; PGM; SNP; wgMLST; surveillance

FF0015

Integrative Biology of a Fungus: Using PacBio® SMRT® Sequencing to Interrogate the Genome, Epigenome, and Transcriptome of Neurospora crassa

Jason Chin (jchin@pacificbiosciences.com)

Pacific Biosciences

PacBio SMRT® Sequencing has the unique ability to directly detect base modifications in addition to the nucleotide sequence of DNA. Because eukaryotes use base modifications to regulate gene expression, the absence or presence of epigenetic events relative to the location of genes is critical to elucidate the function of the modification. Therefore an integrated approach that combines multiple omic-scale assays is necessary to study complex organisms. Here, we present an integrated analysis of three sequencing experiments: 1) DNA sequencing, 2) base-modification detection, and 3) Iso-Seq™ analysis, in *Neurospora crassa*, a filamentous fungus that has been used to make many landmark discoveries in biochemistry and genetics. We show that *de novo* assembly of a new strain yields complete assemblies of entire chromosomes, and additionally contains entire centromeric sequences. Base-modification analyses reveal candidate sites of increased interpulse duration (IPD) ratio, that may signify regions of 5mC, 5hmC, or 6mA base modifications. The Iso-Seq method provides full-length transcript evidence for comprehensive gene annotation, as well as context to the base-modifications in the newly assembled genome. Projects that integrate multiple genome-wide assays could become common practice for identifying genomic elements and understanding their function in new strains and organisms.

Keywords: RNA, Full length RNA sequencing, base modification

FF0016

gEVAL – A Genome Evaluation Browser for Improving Genome Assemblies

William Chow (wchowb@gmail.com), Kim Brugger, Britt Kilian, James Torrance, Eduard Zuiderwijk and Kerstin Howe

Wellcome Trust Sanger Institute, Cambridge, UK

The Genome Reference Consortium's (GRC) role in releasing the major public reference assemblies for human (GRCh38), mouse (GRCm38) and zebrafish (Zv9) drives an on-going need to improve genome curation of these builds. The actual curation, such as identifying and fixing errors by closing gaps, reordering of the clone and/or evaluating haplotypic regions to represent, are conducted under the GRC directive by a partnership of centres which includes the Sanger Institute, the NCBI, the Genome Institute at WashU and the EBI.

In order to aid this curation process, we have developed the genome evaluation browser, gEVAL (geval.sanger.ac.uk). The gEVAL browser provides a one-stop solution to assess the compliance of a given path with a wide range of aligned data viewable in a large browsable genomic region/window. This includes the correct pairing and suitable distance of mapped clone ends, the placement of markers and cDNAs, overlap evaluation, self-comparisons, multi-assembly comparisons, optical map alignments and more. Not only are irregularities highlighted, gEVAL also provides suitable information to resolve them.

gEVAL releases happen frequently in between major assembly releases, allowing the user an up-to-date snapshot of the assembly, which may not be represented in other public databases.

After proving helpful with the curation of the GRC reference genomes, gEVAL browser has been expanded to include assemblies from rat, pig and several helminth organisms.

FF0017

Comparing Strategies for Ultra-Low Input Libraries

Alicia Clum(aclum@lbl.gov), Doina Ciobanu, Alex Copeland, Joanne Lim, Hope Tice, Chia-Lin Wei, Susannah Tringe, Tanja Woyke

Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598

Having enough genomic material for sequencing continues to be a challenge for the community. The ability to make libraries from smaller and smaller amounts of input DNA will open doors for projects previously limited for technical reasons. Several strategies have recently become available. To test which approach is best, serial dilutions from 50ng down to 1pg were done on a mock community, containing 23 bacteria and 3 archaea. These dilutions were used to test three approaches to create libraries for Illumina sequencing. Libraries were made either using Mondrian, a microfluidics system, Multiple Annealing and Looping Based Amplification Cycles (MALBAC), or Nextera, a "tagmentation" approach. We will present the comparison results and the impact of decreasing amounts of input material.

FF0020

Human sequence assembly scaffolding using Irys genome maps

H.Dai[1], A. Pang[1], A. Hastie[1], W. Stedman[1], Z. Dzakula[1], P- Y. Kwok[2], A. Ummat[3], A. Bashir[3], H. Cao[1]
1. BioNano Genomics, San Diego, CA
2. Mount Sinai School of Medicine, New York, NY
3. University of California, San Francisco, CA

Genome mapping technology from BioNano Genomics provides a platform for direct analysis of extremely long genomic DNA (up to multi-megabases) without amplification. *De Novo* assembly of these single molecule images can yield high fidelity contiguous information across long ranges, particularly in highly repeated regions. Resulting genome maps thus greatly complement assemblies using relatively short second- and third-generation sequencing reads.

We have collected 90-fold depth coverage of the human NA12878 sample from the CEU trio and constructed *de novo* consensus genome maps with N50 length of 4.6 Mb. Separately, a Pacific Biosciences sequence based assembly was produced for the same genome with N50 length of 930 kb. By combing data from these two technologies with a custom designed merging pipeline, we were able to generate an assembly having scaffold N50 length of greater than 10 Mb covering more than 2.7 Gb of the human genome. At the same time, we were able to identify potential misassembles by reviewing the inconsistencies between these two complementary technologies.

FF0021

High performance computing platforms for human genome assembly

V.Dergachev, T.Anantharaman, A. Hastie, W.Stedman, F. Trintchouk, M. Saghbini, K. Haden, H. Cao

BioNano Genomics

High quality human genome assembly relies on complex algorithms with large computing power requirements. We will describe the present state and future trends of modern computing platforms in use for this difficult task and discuss strategies of code development that optimally makes use of the compute hardware.

The BioNano Genomics pipeline assembles genome maps from single-molecule images of long DNA molecules (up to multi-megabase lengths) comprising measured distances between fluorescently label 7-base sequence motif. The sophisticated algorithms combine numerical and graph-theoretic techniques to produce de novo assemblies with average contiguous consensus map length up to several megabases.

A recently released Xeon Phi processor can provide 10x more computational capability compared with usual Intel CPU at the cost of increased software scalability requirements.
We will highlight how adapting BioNano software to Xeon Phi platform resulted in gains in computational and power efficiency, as well as improving scalability and speed on regular processors.

FF0022

PacBio sequencing, de *novo* genome assembly, epigenome analysis, and annotation of a novel *Gordonia sp.* isolate

Nico Devitt[1], Barri Herman[2], Randi A. Luchterhand[2], Michael A. Costa[2], Jennifer L. Jacobi[2], Norman G. Lewis[1,2], Laurence B. Davin[2], Callum J. Bell[1]

1 National Center for Genome Resources 2 Washington State University

The genus *Gordonia* is gaining increasing attention due to its relevance in a variety of realms, mainly due to its diverse metabolic activity which has important implications for bioremediation; several *Gordonia* species are able to degrade rubber, organic pollutants and other xenobiotics. It also has notable roles in agriculture and as an opportunistic human pathogen. A filamentous microbe matching the description of a member of the phylogenetic order Actinomycetales was isolated from a soil sample obtained in the Pacific Northwest. The isolate was cultured in modified (10%) Murashige and Skoog media with full Gamborg's vitamins, genomic DNA was isolated with a GenElute™ Bacterial Genomic DNA kit (SIGMA), following the protocol for gram-positive bacteria used in conjunction with Lysozyme (SIGMA), and sequenced on the PacBioRSII platform. The *de novo* genomic assembly yielded 65 scaffolds covering 5,456,845bp. Annotation was performed using the RAST annotation server, yielding 4,906 coding sequences. Comparative analysis of coding sequences confirms the identity of the isolate as a member of the genus *Gordonia*. Base-modification and motif analysis was also performed using the SMRT analysis 2.0 protocols provided by Pacific Biosciences. This epigenomic modification data is combined with the annotations in an effort to further elucidate the regulation of genes of interest, paying special attention to those potentially relevant to bioremediation and pathogenesis.

Key words: Assembly, methylation, annotation, pathogenesis, bioremediation

FF0023

Obtaining Near-Complete Genomes of Novel Bacterial Species from a Sewage System Using Gel Microdroplets and Single Cell Genomics

Armand Dichosa, Karen Davenport, Po-E Li, Cheryl Gleasner, Yuliya Kunde, Kim McMurry, Hajnalka Daligault, Chien-Chi Lo, Ashlynn Daughton, Cliff Han, and Patrick Chain
B-11: Bioenergy and Biome Sciences, Los Alamos National Laboratory, Los Alamos, NM 87545

Most environmental bacteria remain recalcitrant to traditional cultivation methods, likely due to critical factors, yet unknown, related to cell-to-cell signaling and key metabolites lacking in the artificial media. Deep metagenome sequencing and single cell genomics (SCG) offer culture-independent means to obtain genomic data through next-gen sequencing platforms and bioinformatics. However, metagenomics typically result in incomplete genomes, especially from rare and novel species, while SCG by multiple displacement amplification (MDA)[1] provides [mean]_ENREF_2_ENREF_2_ENREF_2 ~40% genome assemblies[2] from a single cell template and has been known to generate chimeras[3]. Recent evidence demonstrated that additional, clonal genomic template greatly improves the chances to obtain complete genomes[4-8]_ENREF_3. However, unless species-specific markers are readily available (e.g., FISH probes, surface fluorescent antibodies) or if the target bacterium is naturally polyploid, researchers are only left with traditional cultivation to generate genomic template from bacterial species that will hopefully grow autonomously.

Building upon our efforts to advance the field of SCG[6-8], our team adapted gel microdroplets (GMDs) to co-cultivate a bacterial consortium originating from a Singapore sewage-wastewater treatment facility. We modified our previous GMD methods for *in vitro* growth[8, 9] to single-capture wastewater bacteria and co-cultivated in 5% of its native environment, and for comparison, in 10% defined R2A medium. In this work, we describe: (a) our cultivation and FACS approaches to monitor bacterial growth in GMDs, and (b) our bioinformatics efforts to assemble near-complete genomes of potentially novel bacterial species isolated from each culture condition, *Thauera* sp. and *Enterobacter* sp., using Illumina-PacBio hybrid and PacBio-only assemblies, respectively. We also demonstrate that bacterial genomes can be sequenced and assembled from clonal microcolonies (i.e., GMDs) amplified by MDA exclusively with the PacBio platform.

This work represents the first, near-*in situ*, high-throughput cultivation approach to assemble genomes of rare, novel bacterial species with GMDs and SCG, and offers a promising methodology to chase after the "unculturable" majority of environmental bacteria.

References:

1. Lasken, R. Single-cell genomic sequencing using multiple displacement amplification. *Current Opinion in Microbiology* **10**, 510 - 516 (2007).
2. Rinke, C. et al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431 - 437 (2013).
3. Lasken, R.S. & Stockwell, T.B. Mechanism of chimera formation during the multiple displacement amplification reaction. *BMC Biotechnology* **7**, 1 - 11 (2007).
4. Seth-Smith, H.M.B. et al. Whole-genome sequences of Chlamydia trachomatis directly from clinical samples without culture. *Genome Research* **23**, 855 - 866 (2013).
5. Woyke, T. et al. One bacterial cell, one complete genome. *PLoS One* **5**, e10314 (2010).
6. Close, D.W. et al. Using phage display selected antibodies to dissect microbiomes for complete de novo genome sequencing of low abundance microbes. *BMC Microbiology* **13**, 1 - 14 (2013).
7. Dichosa, A.E.K. et al. Artificial Polyploidy Improves Bacterial Single Cell Genome Recovery. *PLoS One* **7**, e37387 (2012).
8. Fitzsimons, M.S. et al. Nearly finished genomes produced using gel microdroplet culturing reveals substantial intraspecies diversity within the human microbiome. *Genome Research* **23**, 878 - 888 (2013).
9. Dichosa, A.E.K., Daughton, A.R., Reitenga, K.G., Fitzsimons, M.S. & Han, C.S. Capturing and cultivating single bacterial cells in gel microdroplets to obtain near-complete genomes. *Nature Protocols* **9**, 608 - 621 (2014).

FF0025

Whole genome mapping as a tool to aid genome assembly: Process overview and results from applying it to the assembly of various Helminth worm genomes.

Matthew Dunn[1], Helen Beasley[1], Karen Brooks[1], Michelle Dignam[1], Sarah Nichol[1], Michelle Smith[1], Alan Tracey[1]

[1]Wellcome Trust Sanger Institute

Whole genome mapping is the process whereby long single DNA molecules are linearized, digested with restriction enzymes, viewed, analysed and subsequently assembled into a high definition map, providing a long range structural view of the genome and it's architecture.

Utilising the Argus system from OpGen we have taken two approaches to mapping Helminth genomes. Firstly, for those genomes below 100Mb in size we have generated de novo map assemblies, subsequently aligning the sequence data to the map to provide independent order, orientation and assembly validation of the sequence scaffolds.

The second approach, to improve the assembly of larger Helminth worm genomes (>100Mb), utilizes OpGen's novel Genome Builder process. Genome Builder uses local single molecule assemblies from optical mapping to join sequence contigs together, creating large sequence scaffolds. The resultant local optical map assemblies span several megabases in size, providing previously absent long-range information. We used this information to join sequence contigs and identify regions of misassembly, leading to notable improvement in the current assemblies which further aids the downstream analysis and understanding of these genomes.
These approaches have yielded rapid and significant improvements in the genome assemblies and the methodology and results of both these approaches are outlined across several Helminth projects.

FF0026

PBHoney: Resolving Structural Variation Using Long-Read Sequencing

Adam English, Will Salerno, Min Wang, Christine Beck, Oliver Hampton, Donna Muzny, Jeff Reid, Richard Gibbs

Genomic variation results from numerous biological processes and has been implicated in a variety of diseases. As personal genome sequencing becomes more prevalent, accurate variant identification will further elucidate the nature of human genetic diversity and become increasingly relevant in research and clinical settings. However, the identification of genomic variants via DNA sequencing is limited by both the incomplete information provided by next-gen sequencing (NGS) reads and the nature of the genome itself.

Here we present PBHoney, which identifies structural variants using sequencing data from Pacific Biosciences. Despite the high per-base error rate of PacBio reads, their long lengths and lack of systematic bias allow for highly accurate mapping, even in repetitive regions which are generally enriched for structural variants that are often inaccessible to NGS methods. Were a PacBio read mapped to a reference genome in a region with no structural variation, we would expect to see the read fully mapped (from end to end) with approximately 85% identity. With PBHoney's algorithms, we identify regions where reads map with lower than 85% identity (intra-read discordance) or prematurely stop mapping (interrupted mapping). We then classify these regions as an Insertion, Deletion, or Mismatch in the sample genome relative to the reference.

As a proof of concept, a total of 95,778 PacBio RS filtered subreads were generated, with an average length 6.1 kbp and an N50 length of 8.75 kbp, providing 126.6 average coverage of the 4.6 Mbp E. coli MG1655 Strain K12 genome (GenBank accession U00096.2). Our method identifies a known Transposon Deletion, a tandem duplication and tandem inversion, evidence of the P-element inversion in the e14 prophage, and the presence of the rac-phage in isolate. Furthermore, we sequenced a single human genome to 10X PacBio coverage and identified structural variants with PBHoney. We compared these results with variants called from other structural variant technologies and methods including high-resolution array comparative genomic hybridization and popular techniques used with NGS reads. We show that PBHoney is able to accurately identify variants across a broad spectrum of variant sizes and genomic contexts.

FF0027

Analysis of Mixtures Using Next Generation Sequencing (NGS) of HLA Amplicons: Application to Estimating Proportion of Fetal DNA in Maternal Plasma

Bryan Hoglund[1], Melinda Rastrou[1], Henry Erlich,Ph.D.[1,2], Cherie Holcomb,Ph.D.[1]

[1]Roche Molecular Systems; [2]Children's Hospital & Research Center at Oakland

The analysis of mixtures (samples that contain >one genotype) remains a challenge in both clinical and forensics applications. The massively parallel and clonal aspects of next generation sequencing systems allows for "deep sequencing", the generation of thousands of clonal sequence reads, and "digital analysis", the resolution of the mixture by simply counting the number of reads corresponding to the component genotypes. We have previously developed an HLA typing system based on amplicon sequencing. The highly polymorphic HLA loci represent valuable genetic markers for analyzing mixtures because, typically, the component sequences (majority and minority type) differ by more than a single nucleotide, allowing the distinction of "signal" from "background noise".

For the analysis of mixtures using HLA amplicon sequencing, the HLA typing software (Conexio Genomics) was modified to detect > two alleles for a given sample. Using amplicons for the polymorphic second exon of HLA-DPB1 and -DQB1 loci for the analysis of contrived mixtures of two DNA samples at varying ratios and input, a minority component of 1% could be detected with 1 ng genomic DNA input using the 454 GS Junior and GS FLX instruments. This system was applied to determine the proportion of fetal DNA in maternal plasma, an estimate whose precision is critical in the Non-Invasive Prenatal Diagnosis Test (NIPDT) algorithms designed to detect aneuploidy from NGS data. . Since the cell-free DNA in plasma is short (around 150 bp), the HLA-DPB1 and DQB1 primers were redesigned to amplify and sequence ca. 150 bp second exon fragments on the short read Illumina (MiSeq) as well as on the 454 platforms. The proportion of paternal derived sequence reads in DNA extracted from plasma of a third trimester pregnancy was determined to be 7.4%, 7.5%, and 7.4% in three replicates based on >30,000 reads per sample.

FF0030

LANL Genome Science Program Sequencing Capabilities

Cheryl Gleasner, Yuliya Kunde, Kim McMurry, Momchilo Vuyisich and Tracy Erkkila

Los Alamos National Laboratory, NM

The Genome Science Program of the Bioenergy and Biome Science group (B-11) in Bioscience Division at Los Alamos National Laboratory (LANL) specializes in high-throughput genomics and genome analysis for a variety of internal and collaborative projects. This includes draft sequencing, genome improvement, transcriptome and metagenome sequencing. Sponsored projects are in support of DOD, DHS, DOE and national security missions. We have active projects in many areas, including pathogen biology, biosurveillance, energy and bioremediation. As part of our DTRA-CTR and Department of State collaborations, we have developed a sequencing training program.

The LANL Genome Science Program strives to increase our repertoire and decrease sequencing costs. Current R&D projects include nucleic acid extraction from a wide variety of sources, removal of non-essential RNA, RNA library prep for low input amounts and amplicon sequencing.

The LANL Genome Science Program currently utilizes a number of platforms that include Illumina (HiSeq 2000 and MiSeq), PacBio (RS), and Ion Torrent (PGM and Proton). Depending on the project, data is analyzed in a highly automated fashion with dedicated analysis pipelines, are custom analyzed by bioinformaticists.

FF0036

Taxonomic and functional insight from the metagenome of a pH4 and low temperature thermal spring microbial community in Yellowstone National Park

Xiaoben Jiang[1], Kendra Maas[2], Cristina Takacs-Vesbach[1]

[1]University of New Mexico, [2]University of British Columbia, [1]University of New Mexico,

The taxonomic diversity and metabolic function of the microbial community retrieved from a novel, unexplored pH 4 (55 ºC) thermal spring of Yellowstone National Park (YNP) was analyzed based on 372 Mbp of unassembled metagenomic reads and exhaustive 16S rRNA gene amplicon pyrosequencing. Taxonomic classification of the DNA sequences showed that the community harbored a diverse anoxygenic phototrophic community dominated by members of the Chloroflexi, Bacteroidetes, Proteobacteria and Firmicutes. Functional comparison with other YNP phototrophic metagenomes indicated that the community was enriched in the COG functions related to energy production and conversion, transcription and carbohydrate transport. Analysis of genes involved in nitrogen metabolism revealed the presence of dissimilatory and assimilatory nitrate reduction. Genes involved in sulfur metabolism were mostly related to the reduction of sulfate to adenylylsulfate, sulfite and H_2S. The majority of genes involved in nitrogen and sulfur metabolism were related to the Chloroflexi, Bacteroidetes, Proteobacteria and Firmicutes. Overall, the metagenomic sequencing data outlined the first scenario regarding the microbial assemblage and functions in a pH4 and low temperature YNP thermal spring.

Keywords: Metagenomics, microbial community, pyrosequencing, thermal springs, Yellowstone National Park,

FF0038

New Frontiers of Genome Assembly with SPAdes 3.1

Anton Korobeynikov[1,2], Dmitry Antipov[1], Anton Bankevich[1], Alexey Gurevich[1], Sergey Nurk[1], Andrey D. Prjibelski[1], Yana Safonova[1], Irina Vasilinetc[1], Alla Lapidus[1,3] and Pavel Pevzner[1,4]

1 Algorithmic Biology Laboratory, St. Petersburg Academic University, St. Petersburg, Russia 2 Faculty of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia 3 Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia 4 Department of Computer Science and Engineering, University of California, San Diego, USA

Despite all the efforts high quality genome assembly is a complex task that so far remains unsolved. It is well known that majority of problems caused by repeats present in all genomes of any nature. The usage of multiple methods of genomic DNA isolation, different sequencing technologies and different types of genomic libraries for research projects introduces additional levels of complication to the genome assembly. The assembler tool SPAdes was originally developed at the St. Petersburg Academic University for the purpose of overcoming the complications associated with single-cell microbial data (uneven coverage and increased level of chimerical reads). The tool was able to successfully resolve these issues for Illumina reads and was recognized by the scientific community as one of the best assemblers working with both isolates and single-cell data.

Current SPAdes 3.0 version of the assembler provides an ability to work with different combinations of sequencing platforms including Illumina, Ion Torrent, PacBio, Sanger using both paired-end and mate-pair libraries of different insert sizes. Here we present SPAdes

3.1 - further development of SPAdes, which resolves many scaling issues with respect to running time and RAM consumption as well as improved support for IonTorrent data, updated algorithms for scaffolding and repeat resolution, and an approach for mate-pair only assembly from the reads produced by novel Illumina NexteraMP protocol.

Keywords: de novo assembly, hybrid assembly, repeat resolution, scaffolding, mate pairs

Also to be presented as talk.

FF0039

Whole genome sequence comparison of ten diagnostic brucellaphages propagated on alternate bacterial hosts

Adam Kotorashvili[2], Ekaterine Tevdoradze[1]*, Jason Farlow[3,4]*, Natia Skhirtladze[1], Irina Antadze[1], Sophio Gunia[1], Nana Balarjishvili[1], Leila Kvachadze[1]and Mzia Kutateladze[1]

1. George Eliava Institute for Bacteriophages, Microbiology and Virology, Tbilisi, Georgia 2. Lugar Center for Public Health Research at National Center for Disease Control, Tbilisi Georgia 3. Academic Engagement Program (AEP) Pennsylvania State University, USA 4. Farlow Scientific Consulting Company, LLC, Utah USA

To elucidate fine-scale trends in the genetic structure of brucellaphages we performed whole genome sequencing of ten diagnostic phages following propagation on alternate host strains of B. abortus including the phages Tb, 1066, 281, 02, 177, 110, 11sa, 544, 141, and V. All phages in this analysis lacked the two major deletions present in the other non-B. abortus phage. Whole genome sequence analysis revealed substantial genetic homogeneity consistent with previously published data; however, several mutations emerged in select genes following propagation from host strain 141 to strain S19. Genomic alterations were observed in similar genes across multiple phages, and predominantly occurred at identical sites across separate phage lineages. The Tb phage displayed phenotypic and genetic differences that may be mediated by alternate biological attributes across distinct though closely related bacterial hosts. Genomic alterations in Tb following propagation from 141 to S19 included SNPs predominantly in ORF 21 (positions 15576 and 15578) and the phage tail collar gene (positions 2185 and 22178). Furthermore, positive selection was detected in the tail collar protein gene consistent with previous data from diverse brucellaphages across the genus. We also identified a Staphlothermusmarinus F1-like CRISPR spacer and sequences orthologous to both prophage antirepressors of Brucella spp. And intergenic sequences encoded by Ochrobacterium anthropi. Overall, these data provide insight into common fine-scale differences emerging within brucellaphage genomes during propagation across alternate B. abortus host strains and indicates candidate loci for future dissection of brucellaphage host range determinants.

FF0040

Chromosome-scale assembly of the lettuce (*Lactuca sativa*) genome using genotyping by shotgun sequencing (GBSS)

Alexander Kozik, Lutz Froenicke, Dean Lavelle, Maria Jose Truco, Huaqin Xu, Sebastian Reyes Chin-Wo, and Richard Michelmore

Genome Center, University of California, Davis CA, 95616, US

Next Generation Sequencing (NGS)-based assemblies of moderately large genomes [1-3 Gb] often consist of thousands of scaffolds. The combination of NGS and segregation analysis allows the generation of saturated, ultra-high density genetic maps, which can be employed to improve the contiguity of genome assemblies. The genome of the cultivated lettuce, *Lactuca sativa*, cv. Salinas has been sequenced in collaboration with the BGI supported by a consortium of ten breeding companies. The SOAPdenovo assembly of the *L. sativa* genome (2.7 Gb) consisted of 21,500 scaffolds covering 2.38 Gb [http://lgr.genomecenter.ucdavis.edu/]. We also sequenced *L. serriola*, a close relative of *L. sativa* and its likely progenitor. In addition, 98 Recombinant Inbred Lines (RILs) from a cross between *L. sativa* x *L. serriola* genotypes was sequenced at 1x coverage. We then used Genotyping by Shotgun Sequencing (GBSS) to generate a high resolution map of the Lactuca genome. Comparison of the sequences from *L. sativa*, *L. serriola* and the RILs using CLC Genomics Workbench revealed more than 11 million reliable high-quality SNPs. Custom scripts were used to assign haplotypes to scaffolds and to identify chimeric misassemblies. The latter were corrected by breakage at the junction point. GBSS and segregation analysis of the RILs allowed the chromosomal assignment of more than ten thousand scaffolds and their ordering into genetic bins, covering 2.3 Gb of the lettuce genome. The resulting comprehensive chromosome assemblies are facilitating the identification and functional analysis of candidate genes for a broad spectrum of domestication and agriculturally-important phenotypes.

Keywords: variant analysis, genotyping by sequencing, assembly validation, scaffold ordering

FF0042

Assembly of Repeat Content Using Next Generation Sequencing Data

K. LaButti (klabutti@lbl.gov) , A. Copeland, A. Kuo

Lawrence Berkeley National Laboratory, US Department of Energy Joint Genome Institute

Short read technology has enabled efficient and high quality assembly of prokaryotic genomes. However, highly repetitive organisms still pose a challenge for short read assembly, and typically only unique regions and repeat regions shorter than the read length, can be accurately assembled. Recently, we have been investigating the use of Pacific Biosciences long reads for de novo fungal assembly. We will present an assessment of the quality and degree of repeat reconstruction possible in a fungal genome using long read technology. We will also compare differences in assembly of repeat content using short read and long read technology.

FF0043

Fast and comprehensive analysis of large structural variants in human genomes

E. Lam, A. Hastie, W. Andrews, H. Dai, M. Austin, F. Trintchouk, M. Saghbini, T. Anantharaman, K. Haden, H. Cao

BioNano Genomics, San Diego, California, US

Irys genome mapping represents a recent single-molecule platform complementary to short-read sequencing for genome assembly and structural variation analysis. Extremely long molecules of hundreds of kilobases fluorescently labeled at specific sequence motifs span across and enable direct interrogation of structural variants.

The short turnaround time to comprehensively analyze a human genome has given us the ability to rapidly analyze multiple genomes and perform cross-sample comparison to identify variation. To date, we have *de novo* assembled more than 20 normal and diseased human genomes and analyzed their structural variation contents. Genome map assemblies cover the majority of non-N base portions of the genome but also extend into subcentromeric and subtelomeric regions of the genome. We have expanded our analysis pipeline to include detection and validation of inversions and translocations in addition to deletions and insertions. We detected hundreds of large structural variants per genome and haplotype differences. Furthermore, we constructed highly accurate copy number profiles that are free from amplification bias and are particularly informative for analysis of cancer genomes.

FF0045

In silico **analysis of whole genome sequences for virulence genes profiles of Shiga toxin-producing *Escherichia coli* (STEC) O26, O45, O103, O111, O121, O145 and O157 from outbreaks and apparently sporadic infections in the United States.**

Rebecca L. Lindsey, Nancy A Strockbine, Eija Trees, Haley Martin, Devon Stripling, Patricia Lafon, Ashley Sabol, Heather Carleton, and Peter Gerner-Smidt.

Centers for Disease Control and Prevention, Atlanta, GA

Shiga toxin-producing *Escherichia coli* (STEC) belonging to serogroups O26, O45, O103, O111, O121, O145 and O157 are the most frequently isolated STEC from outbreaks and sporadic infections in the United States. STEC carry at least one Shiga toxin gene and frequently carry intimin and hemolysin genes. We analyzed whole genome sequence (WGS) to determine the virulence gene profiles of a collection of isolates from outbreaks and apparently sporadic infections and compared the results to conventional laboratory testing.

125 STEC isolates belonging to serogroups O26 (25), O45 (5), O103 (5), O111 (25), O121 (25), O145 (15) and O157 (25) were sequenced with Illumina Hiseq and assembled into an average of 200 contigs per isolate. Five isolates lacking the above STEC virulence genes were included as negative controls. Analysis was completed using CLC Genomics Workbench 6.5.1. Previously published primers were tested and the most accurate sequences were selected for each gene. The best *in silico* conditions included a minimum number of base pairs required for a match to be 18 and at least 8 consecutive base pairs were required at the 3'end. When *in silico* results were compared to laboratory results; 98% of *stx1*, 93% of *stx2*, 100% of *eae* and 95% of *hylA* matched. Additionally *in silico* subtyping of the above genes was completed. Subtyping of these genes is currently not performed in the laboratory due to time and monetary constraints.

Sequence based detection of genes for diagnostics and subtyping of STEC are the first step in the development of more comprehensive tools like BioNumerics -based whole genome multi-locus sequence typing (wgMLST). It is important that future bioinformatic tools are backwards compatible with established methods for gene detection as laboratories move from conventional to sequence based methods for detection of virulence genes.

Keywords: Escherichia coli, STEC, virulence, CLC

FF0046

FaQCs: Quality Control Of Next Generation Sequencing Data

Chien-Chi Lo and Patrick Chain

Bioenergy and Biome Sciences, Los Alamos National Laboratory

Next generation sequencing (NGS) technologies that parallelize the sequencing process and produce thousands to millions, or even hundreds of millions of sequences in a single sequencing run, has revolutionized genomic and genetic research. The Illumina® HiSeq sequencing platform can now produce upward of six billion sequencing reads (at 100 bp each is 600 Gb) in a single 11---day run. Another recent platform, the Life Technologies® Ion Proton is projected to have a throughput of 120 Gb in an 8---hour run. Because of the vagaries of the sequencing chemistry, the experimental processing, machine failure, and so on, the quality of sequencing reads is never perfect, and often declines as the read is extended. These errors invariably affect downstream analysis and applications and should be identified early on to mitigate any unforeseen effects. Here we present a rapid parallelized program that can process such large volumes of data and which improves upon previous solutions to monitor the quality and remove low quality data from sequencing runs. This tool combines several features of currently available applications into a single, user--- friendly process, and includes additional unique capabilities such as filtering the PhiX control sequences, conversion of FASTQ formats, and multi---threading. Both the speed of processing and the required memory footprint of storing all required information have been optimized and can be further refined by the user via parallel processing. Finally, the original data and trimmed summaries are reported within a variety of graphics and reports, providing a simple way to do data quality control and assurance.

FF0049

Studying variation and experimental noise in RNA-seq data

Punita Manga[1,2], Dawn M. Klingeman[2,3], Tse-Yuan S. Lu[3], Tonia L. Mehlhorn[4], Dale A. Pelletier[3], Loren J. Hauser[1,3], Charlotte M. Wilson[1,2,3], and Steven D. Brown[1,2,3,§].

[1]Graduate School of Genome Science and Technology, University of Tennessee; [2] BioEnergy Science Center, Oak Ridge National Laboratory; [3] Biosciences Division, Oak Ridge National Laboratory,[4] Environmental Sciences Division, Oak Ridge National Laboratory

Continued improvements and next-generation sequencing cost reductions have made RNA-seq a popular choice for gene expression studies. RNA-seq is revolutionizing the fields of genomics and transcriptomics by enabling a wide range of applications such as strand specific expression detection, alternative splicing isoforms, SNP detection, genome guided or *de novo* transcript assemblies. It also enables detection of weakly expressed genes and high genome coverage adds to the attraction as an alternative to microarrays. However, the field of RNA-Seq analysis is still evolving. This study aimed to identify the technical and experimental variations in RNA-seq datasets in order to minimize unwanted variation in transcriptomic studies.

RNA-seq data for *Bacillus thuringienesis* strain CT43 and 10792 were obtained using Illumina High-Seq2000 and subjected to quality assessments, normalization (KDMM, TMM, UQS, RPM and RPKM), distribution analysis and statistical tests to analyze experimental noise. Data for four biological replicates were obtained and genome coverage ranged from 85-300X. The number of significant genes varied with different normalization methods. The number of differentially expressed genes (5% False Discovery Rate and two-fold change) for strain 10792 was 459, 483, 460, 409 and 194 for KDMM, UQS, RPKM, RPM and TMM normalization methods, respectively. Variation sources included LB medium lots/water sources and date of culture. Genes related to iron acquisition and metabolism were consistently differentially expressed for both strains grown in one medium lot compared to another. We hypothesized there were differing amounts of iron in the LB medium prepared from the two different lots, which was confirmed by elemental analysis.

RNA-seq is a highly sensitive and powerful technique, which can with the appropriate normalization methods and statistical models identify noise from various sources and be used for differential expression studies. Experimental noise can be controlled and minimized with appropriate design of experiment, biological replicates, normalization and during statistical testing.

Keywords: RNA-seq, *Bacillus thuringienesis,* Normalization, Experimental noise

Acknowledgments: This work is sponsored by the Laboratory Directed Research and Development Program of ORNL. Stanton L. Martin, Tzu-Ming Chu, Russ Wolfinger (SAS Institute).

FF0050

Resequencing Workflows: Variant Calling in Deep and Shallow Coverage NGS Samples

Marta Matvienko[1], and Alexander Kozik[2]

[1]CLC bio, a Qiagen Company, [2]Genome Center, University of California – Davis

Genotyping-by-sequencing is becoming widespread in breeding research. Efficient data analysis pipelines require that bioinformatics tools are accurate and easy-to-use, and can be optimized and automated. We describe methodological procedures for calling variants in parental libraries and Recombinant Inbred Lines (RILs) after NGS sequencing. The parental lines (*Lactuca sativa* and *Lactuca serriola*) were sequenced at 12x (deep) coverage, and the 99 distinct RILs at 1x (shallow) coverage.

The sequencing data used in this analysis is generated and owned by Richard Michelmore lab at UC Davis Genome Center. The sequencing was done using HiSeq2000, and the data analysis was performed with CLC Genomics Server and Workbench, CLC bio, http://www.clcbio.com/ .

The parental libraries were used to create the high quality (HQ) variations track. The reads from *L. sativa* and *L. serriola* were mapped separately to *L. sativa* genomic scaffolds using stringent conditions. The local realignment tool was applied to improve the alignment of individual reads in the presence of indels and unaligned ends. The variations were called using the Probabilistic Variant Detection tool (minimum coverage=3, min variant probability=99%). To filter for HQ nucleotide variants, additional requirements for coverage and frequency in control, zygosity, type, and others were applied. This produced the HQ reference variant track.

For calling variants in shallow sequenced RILs, the same mapping parameters were used, but the stringency of parameters in the variant caller was reduced (minimum coverage=1; variant probability=50%). To select the reliable variants in the RIL libraries, these tracks were compared with the HQ reference variant track using the “Filter against Known Variants” tool. The resulting mapping and variant tracks can be integrated into a project (track list), and inspected in the graphical viewer.

All steps described above can be compiled into workflows using the graphical workflow creation tool.

FF0052

Detection and comparison of large clinically relevant insertions and deletions

Rob Mervis, Anne-Mette Hein[1] , Patrick Dekker[1], Anika Joecker[1], Cecilie Boysen[1], Naomi Thomson[1], Bodil Øster[1], Anne Arens[1,]Bjarne Knudsen[1], Thomas Knudsen[1], Roald Forsberg[1]

[1]CLC bio, a QIAGEN Company, Silkeborgvej 2, 8000 Aarhus, Denmark

Large somatic insertions and deletions in tumor samples are often of significant clinical impact. Many of them are cancer driver mutations and play an important role in drug treatment[1].

Detecting the accurate breakpoints of larger insertions and deletions is often problematic and inaccurate. Furthermore, due to ambiguous positioning of them in the human genome, they are hard to compare with known deletions and insertions in publicly available databases.

Here we present a complete analysis workflow to identify, filter and annotate large insertions and deletions with information from publicly available databases such as COSMIC. We apply our pipeline on publicly available tumor/normal matched pair data from a patient with Massive Acinic Cell Carcinoma (A. Nichols et al. (2013) Case Report in Oncological Medicine, Article ID 270362), which has not been investigated for larger insertions and deletions before. We show that we can identify insertions and deletions, which should be reported as part of the list of somatic variants, the authors have presented. Moreover, we will point out potential challenges and how to solve them while comparing own results to data in databases.

1 Katerina Politi and Thomas J. Lynch1 (2012). Two Sides of the Same Coin: *EGFR* Exon 19 Deletions and Insertions in Lung Cancer. *Cancer Res March 15, 2012 18;* 1490

FF0054

Microbial Genome Assembly and Finishing: Automated and Manual Tools

David Michaels, Marta Matvienko, Rob Mervis, Poul Liboriussen, Peder Roed Lindholm Nielsen, Jesper Jakobsen, Steffen Mikkelsen, Henrik Sandmann, Søren Mønsted, Jannick Dyrløv Bendtsen, Martin Simonsen

CLC bio, a Qiagen Company

De novo genome assembly and genome finishing is becoming increasingly important in the sequencing projects. The hybrid data comprised of short and long reads can complement each other to generate high quality assemblies.

Here we demonstrate how an assembly of a microbial genome can be optimized using the CLC Microbial Genome Finishing Module. The module is a collection of tools for identifying, visualizing and solving problems in genome assemblies from short NGS reads, long NGS reads, and Sanger reads. The publicly available MiSeq and PacBio data for *E. coli* were used to optimize the parameters for genome assembly and finishing. The contigs were assembled from MiSeq data in the Genomics Workbench. To evaluate and improve the assembly, we used the Microbial Genome Finishing Module.

The contigs were analyzed using the Analyze Contigs tool. This identified and annotated the problematic regions that needed further attention. Those were the regions with low, high, single-stranded, unstable coverage, and regions with unaligned read ends. After manual inspecting and editing mappings in these areas, we used the automated Join Contigs tool with the option to join contigs with uncorrected PacBio reads. This step dramatically reduced the number of contigs , and increased the N50 length. Besides using long reads, the Join Contigs tool provides an automated way of joining contigs based on other types of analyses:

> Paired reads that are mapped to different contigs are used to identify the neighboring contigs, the distance between them, and orientation;
> BLAST alignment of contigs against each other;
> When a closely related reference is available, the contigs are joined using the BLAST alignments to the reference;

To close the remaining gaps in the genome, the primers for the ends of all contigs, and for the low coverage regions, could be designed using the automated primer design tool.

FF0055A

High Efficiency Long Insert Mate Pair Library Preparation for NGS Platforms

Scott Monsma (smonsma@lucigen.com), David Mead, Svetlana Jasinovica, Erin Ferguson, and Michael J. Lodes

Lucigen Corporation, 2905 Parmenter Street, Middleton, WI 53562, USA

Next generation DNA sequencing (NGS) instruments produce gigabases per run, but the short read lengths and small size of sequenced fragments result in gaps, misassembled contigs, collapsed repeats and missing sequences, leaving these regions to be finished manually, if at all. A technology that provides long range sequence linkage from short reads is needed for accurate, economical *de novo* assembly of genomes. We have developed a >90% efficient mate pair library construction technology that incorporates Chimera Codes™ to distinguish true mate pairs from false junctions. NxMate™ NGS libraries were constructed using a reference *E. coli* strain, *Thermus aquaticus* and two repeat rich mouse BACs. Without mate pair libraries the BAC and genome assemblies contained numerous unordered contigs. The addition of >90% efficient NxMate data allowed accurate de novo assembly and closing of the BACs and genomes. For comparison, several Illumina Nextera mate pair libraries were constructed and yielded <5% mate pair data. Ion Torrent "mate pair reads without paired end sequencing" permits economical sequence assembly of BAC clones and small genomes, while NxMate libraries for the Illumina platform allows for long range paired end sequencing and assembly of larger genomes.

FF0055B

40 kb Mate Pair NGS Library Technology

David Mead[1], Svetlana Jasinovica[1], Megan Wagner[1], Amanda Krerowicz[1], Ronald Godiska[1], Cheng-Cang Wu[1], Amy Hin-Yan Tong[2], Si Lok[2], Matthew E. Hudson[3,4], Therese Mitros[3,5], Daniel S. Rokhsar[3,5,6], Kankshita Swaminathan[3,4] and Stephen P. Moose[3,4]

Presented by *<u>Scott Monsma</u>*

1. Lucigen Corporation, Middleton, WI 53562 USA . 2. School of Life Sciences, The Chinese University of Hong Kong, Hong Kong SAR, China, 3. Energy Biosciences Institute, 4. University of Illinois, 5. University of California Berkeley, 6. DOE Joint Genome Institute.

Next generation sequencing (NGS) technologies can rapidly and economically produce a draft genome of an organism de novo. However, the quality of the draft data is seldom more than 80% complete with >10e6 contigs for large genomes, which is insufficient for many applications. Most contigs begin and end with a repeat with existing library construction technologies. Sequence data that is closer to 95% assembled with unambiguous order and placement of genes would have the greatest utility for scientific and commercial research. New molecular tools that bridge the gaps between massively parallel short read sequencing technologies (35-1,500 bases) and large scaffolds (>100,000 bases) are needed to accurately assemble complex repeat rich genomes. We have successfully produced 40 kb mate pair NGS libraries by designing and constructing a novel fosmid system. Our results show that ~70% mate pair sequences can be obtained from a reference human genome. Assembly of the 2.5 Gb *Miscanthus sinensis* genome was increased from N50 40 kb to 102 kb using 960,000 clones from a 40 kb mate pair library.

FF0056

KAPA Hyper Prep: A next-generation kit for fast and efficient library construction from challenging DNA samples

Joseph Musmacker[1], Victoria van Kets[1], Maryke Appel[1], Jacob Kitzman[2], Matthew Snyder[2], Jay Shendure[2], Phillip Gray[3], Pei-Fang Tsai[3], Jia Yan Hoo[3], Eric van der Walt[1], John Foskett[1] and Paul McEwan[1]

[1]KAPA Biosystems, Inc., [2]Genome Sciences, University of Washington, [3]Ambry Genetics

Kapa Biosystems has previously developed a robust and versatile DNA library preparation kit, which is currently regarded as the best sample preparation solution for Illumina sequencing. A combination of ultra-pure, optimally-formulated enzymes and a highly-optimized “with-bead” protocol enables the construction of high-quality libraries from challenging samples such as FFPE, ChIP and cell-free DNA.
Rapid growth in sequencing capacity, falling costs and the implementation of next-generation sequencing (NGS) in clinical and diagnostic settings are driving a relentless demand for NGS library construction solutions that are faster, simpler, more cost-effective, and easier to scale. This is evidenced by the growing interest in streamlined, “single-tube” ligation-mediated library construction workflows, and transposase-mediated “tagmentation” technology. Typically, single-tube ligation- or tagmentation-based library construction methods represent considerable compromises in performance: they convert less input DNA into sequenceable molecules, resulting in libraries with lower diversity and reduced coverage uniformity; and are more sensitive to the amount and/or quality of input DNA. This presents a particular challenge for projects involving challenging samples of variable input and quality.

To address these requirements, Kapa Biosystems has developed a streamlined library construction method that enables fast and highly efficient library construction from a range of sample types and inputs. The KAPA Hyper Prep Kit performs as well or better than the existing KAPA Library Preparation Kit with high-quality DNA across an input range of 10 ng – 1 µg. The novel single-tube chemistry offers specific benefits for library construction from FFPE and low-input (<10 ng) samples such as cell-free DNA.

FF0059

Haplotyping Assembly Refinement and Improvement II

Olsen, C. (Christian@biomatters.com)[1], Qaadri, K.[1], Shearman, H.[2], Miller, H.[2], Ammundsen, B.[2] Hsiau, T.[3], Rudenko, G.[3], Huynh, T.[3], Somanchi, A.[3], Zhao, X.[3], Brubaker, S.[3]

[1] Biomatters, Inc. 60 Park Place, Suite 2100 Newark, NJ 07102 [2] Biomatters, Ltd. Level 2, 76 Anzac Ave. Auckland 1010 New Zealand [3] Solazyme, Inc. 225 Gateway Blvd. South San Francisco, California 94080

Haplotyping is the assignment of polymorphisms to the correct allele(s) in a diploid or polyploid organism. Haplotyping is an especially complex problem that has been little addressed by current technologies. In particular, current Next-Generation Sequencing (NGS) technologies that use short reads make it difficult to haplotype over long distances. A related complex problem is that of diploid or polyploid assembly, which requires the accurate extension of reads to assemble distinct haplotypes for each allele. The aim of the haplotype assembly 'problem' is to reconstruct the two haplotypes using a mix of sequenced fragments from the two chromosomes. This problem has been shown to be computationally difficult for automation and optimization. Current sequence assemblers are designed to try and collapse/merge two or more alleles that may be present in an organism. Applying these assemblers to sequences from diploid or highly polymorphic organisms leads to many problems and mis-assemblies, which have been well documented. However, an effective solution to this problem has not yet been presented. With this talk/poster we present a semi-automated method for assembling and haplotyping a very high-GC, polymorphic, diploid organism.

FF0060

Comparison between known *Salmonella enterica* serotypes and unique variants as determined by intergenic sequence secondary structure (poster)

Olsen, C.[1], Qaadri, K.[1], Shearman, H.[2], Miller, H.[2]

[1] Biomatters, Inc. 60 Park Place, Suite 2100 Newark, NJ 07102 [2] Biomatters, Ltd. Level 2, 76 Anzac Ave. Auckland 1010 New Zealand

PCR-intergenic sequence ribotyping (ISR), a typing method based on polymorphisms in the 5S-23S intergenic spacer regions. This typing is used for distinguishing individual strains and is required for epidemiological and environmental studies. Ribotyping is generally considered the best method for typing. However, it fails to account for sequence diversity which, might exist in intergenic 5S-23S rRNA bacterial operon spacer regions within and among strains of an organism. Determining relatedness of unknown strains using phylogenetic trees of ISRs is problematic because the sequences cluster by size. ISR regions may have unique structural features that help establish evolutionary relationships. Furthermore, DNA folding prediction may provide an additional method for understanding genetic relatedness between dkgB-linked ISRs generated from *Salmonella enterica* subsp I. With this talk/poster we present a workflow for determining single nucleotide polymorphisms and comparison of secondary structures occurring in the ribosomal gene region and flanking sequences of *Salmonella spp*. This method can be used to assign serotype to *Salmonella enterica* by ISR and allows for efficient categorization of unknown Salmonella isolates. This simple, cost effective, and efficient PCR-based technique can be used to identify strain isolates which will allow a greater measure of understanding virulence mechanisms and ecological niches.

FF0061

Viral Metagenome Pipeline

Olsen, C.*[1], Qaadri, K.[1], Shearman, H.[2], Miller, H.[2], Ammundsen, B.[2]

[1] Biomatters, Inc. 60 Park Place, Suite 2100 Newark, NJ 07102 [2] Biomatters, Ltd. Level 2, 76 Anzac Ave. Auckland 1010 New Zealand

Metagenomic analysis is quickly gaining momentum as a well-suited technique to provide a detailed understanding into the composition and activity of bacterial and viral communities i.e. “microbiome”, “virome”. Understanding the role of the virome in health and disease requires a deeper understanding of their composition and dynamics under various environmental conditions within the human gut and other body sites. As the number of sequenced viromes increase, larger genomic fragments are resolved by assembling the large amount of sequence data generated for each metagenome. In this talk, we present a pipeline for virome analysis consisting of data preprocessing, assembly, annotation, and comparison. The metagenomic pipeline is able to utilize large data-sets comprising of viromes made of thousands of large genomic contigs. This pipeline can be used to analyze two types of datasets: (i) viromes composed of raw reads, mostly generated using 454 pyrosequencing technology and (ii) viromes assembled into contigs, a strategy made possible with datasets sequenced with either/both pyrosequencing or Illumina sequencing technologies. Users are able to explore and analyze viromes composed of raw reads or assembled fragments using Geneious R7, a user-friendly interface to explore any kind of virome and enables virologists to make the most of their metagenomics next generation sequence data.

Also to be presented as talk.

FF0066

Salmonella enterica serovar Enteritidis outbreak in an institutional setting – what can genomic sequence data tell us?

Ashley Sabol[1], Heather Carleton[1], Valarie Devlin[2], Patricia Lafon[1], Efrain M. Ribot[1], Bradley Tompkins[2], Keeley Weening[2], and Eija Trees[1]

[1] Centers for Disease Control and Prevention, Atlanta, GA [2] Vermont Department of Health, Burlington, VT

In December 2013, a case of *Salmonella enterica* serovar Enteritidis was reported at a nursing home. Over the course of two months, four additional cases were identified from the same facility, one of whom became an asymptomatic long term carrier. Four closely related PFGE pattern combinations were detected and no common exposure was identified. Since it was not clear whether the outbreak strain was evolving during the outbreak or whether different exposures were involved, genomic sequencing was performed to determine diversity among the carrier isolates and other clinical and environmental isolates collected at the facility.

Eighteen isolates, including 11 consecutive isolates from the carrier, were sequenced with the Illumina Miseq using 2x150bp chemistry. Assemblies and hqSNP calls (20x coverage, 95% frequency) were done using CLCBio Genomics Workbench.
Three lineages were detected among the carrier isolates. Lineage A included the original isolate and eight additional isolates which differed from the original one by a median of 48 hqSNPs (range: 22-71). Lineage B was recovered 10 days after the original lineage A isolate and differed by 299 hqSNPs, and lineage C was collected two months after A and differed from A and B by 121 and 224 hqSNPs, respectively. Three of the four other cases from the same facility and the two environmental isolates were within 39 or fewer hqSNPs from the original lineage A isolate.

In conclusion, the main outbreak strain (lineage A) evolved significantly (up to 71 hqSNPs) within the carrier over two months. The two divergent lineages (B and C) more likely represent multiple exposures or the presence of multiple strains in the initial exposure rather than evolution from the main outbreak strain. These results highlight the challenges of using hqSNP analysis to clarify epidemiological relationships in complex settings over an extended timeframe.

Keywords: *Salmonella*, Enteritidis, hqSNP, sequencing, carrier

FF0068

The Exome Coverage and Identification (ExCID) Report: a gene survey tool for clinical sequencing applications

Rashesh Sanghvi[1], Christian Buhay[1], Qiaoyan Wang[1], Harsha Doddapaneni[1], Yi Han[1], Huyen Dinh[1], Eric Boerwinkle[1,2], Donna M. Muzny[1] and Richard A. Gibbs[1]

[1]Baylor College of Medicine, Houston, TX 77030. [2]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX 77225

The Exome Coverage and Identification (ExCID) Report was developed at the Human Genome Sequencing Center (HGSC) to calculate gene transcript and exon sequence coverage for every VCRome Whole Exome sequencing (WES) sample. Since its implementation in March 2013, the report has been generated on more than 13,000 WES research samples for the HGSC and over 1500 WES clinical samples for the Whole Genome Laboratory (WGL) at Baylor College of Medicine.

ExCID assesses target sequencing coverage, annotates targets with gene, transcript and exon information, and reports intervals below 20X. For a pilot survey, coverage data was examined from 34 WES samples in the WGL. Results from aggregate data reveal that by using BCM-HGSC VCRome capture reagent, we fully cover (at ≥20X) over 75% of the medically interesting genes listed in COSMIC, HGMD, GeneTests, and ACMG Incidental Findings lists. The analysis also identified gene regions that were routinely below the 20X threshold.

Clinical sequencing can be divided in two strategies: Whole Genome Shotgun (WGS) and Targeted Sequencing. Each strategy complements different analyses from the identification of complex structural variants to the discovery of rare mutants via deep sequencing. Regardless of the method used, regions of the genome remain difficult to sequence with adequate read depth. To address poorly covered regions in both strategies, ExCID output was generated on 9 WGS (30X) and 9 WES (100X) to discover and aggregate coding regions below 20X in genomes and exomes. The data implies that there are regions that are poorly covered irrespective of the sequencing strategy, as well as regions unique to each strategy. Ongoing analyses include the aggregation and characterization of poorly covered regions in 100 WGS and WES samples. Results will provide insight regarding areas in Human Genome that would require special consideration in the development of future clinical sequencing strategies.

Keywords: Bionformatics, Genomics, Tools, Pipeline

FF0070

Improved analysis of RNA, FFPE RNA and microRNA using the Fragment Analyzer™.

Steve Siembieda, Kit-Sum Wong and Jeremy Kenseth

Advanced Analytical

Next Gen Sequencing (NGS) of RNA is growing quickly and may soon be the method of choice to analyze RNA sequence composition, expression levels, splice sites and variations. Inherent to its analysis, by any molecular method, is an understanding of both the quality and quantity of this extracted nucleic acid. Today, high quality RNA can be easily extracted from a few to many thousand freshly prepared cells. New technologies such as the Fluidigm C_1™ prep system can be used to isolate, extract and convert RNA from single cells into sequencable libraries.

For poor quality RNA samples, like formalin-fixed paraffin-embedded (FFPE) or poorly frozen tissues, RNA quality can suffer considerably. Though sequencable RNA can be obtained and converted into library preparation, care must be taken to understand the initial RNA quality in order to generate reliable results. Additionally, microRNA analysis by NGS sequencing is gaining traction; however, instruments capable of analyzing the quality of these RNA species are limited.

Tested and confirmed to analyze all types of RNA, the *Fragment Analyzer*™ offers both hardware and software features that can improve laboratory workflow by streamlining experimental setup, analyzing samples in parallel and providing intuitive customizable data analysis. For single cell transcriptomics, the *Fragment Analyzer*™ is an ideal tool for verifying single cell capture and amplification when incorporated into the C_1™ sequencing workflow. For total RNA quality, the *Fragment Analyzer*™ RNA Quality Number (RQN) has been validated across a variety of sample types and quality to be equivalent to the traditional integrity score value. For FFPE RNA, recent methods developed by both Illumina and Advanced Analytical show how a DV_{200} metric (i.e. the percentage of RNA >200 nucleotides) can be a better predictor of sequencing outcome than the traditional integrity score value. In the case of microRNA, the *Fragment Analyzer*™ can accurately assess both the quality and quantity of these RNA species in a single separation with high resolution and sensitivity. Data regarding these applications will be presented.

FF0072

Whole Genome de novo sequencing with long read technology

Nick Sisneros (nsisneros@pacificbiosciences.com)

Pacific Biosciences

Significant advances in single-molecule-sequencing read lengths enable highly contiguous de novo assembly of genomes from microbes (Mbase scale) up to complex eukaryotes (multi-Gbase scale). Improvements in raw-read error correction, assembly algorithms, and consensus polishing of Pacific Biosciences' long-read data have led to high-quality assemblies that rival Sanger "clone-by-clone" sequencing efforts. As a demonstration, we have whole-genome shotgun sequenced model genomes such as *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Spinacia oleracea*. We present here the guidelines and practices for achieving optimal results for long read sequencing these genomes using the PacBio® RS II system. We further describe SMRTbell™ library preparation methods and gel purification methods to obtain suitably long insert libraries to obtain maximal sequencing results whereby average sequencing read lengths of 9 kbases can be achieved with extreme reads in the distribution of > 40kbases. The benefits of long reads are demonstrated by the highly contiguous assemblies with contig N50s of over 5Mbases as compared to similar assemblies using next generation short read approaches. Finally, guidelines will be presented for planning out projects for the de novo assembly of large genomes.

Keywords: Best practices, whole genome sequencing

FF0073

Increasing the value of genomics and proteomics data with educational applications

Todd M. Smith and Sandra G. Porter

Digital World Biology, Seattle WA.

Modern data collection technologies are creating enormous data resources that are underutilized in science education and research. We often present graphs showing exponential growth of one kind of database or another, but how often do we stop to ask the questions: how many people are using these data? And how are they using it? Without efforts toward improving general use, data resources will not be valued, which can negatively impact the value of future data collection endeavors.

New approaches are required to increase the overall understanding of how the data can be utilized. The approaches also need to be combined with user-friendly tools and content that demonstrates specific applications with interesting stories. Digital World Biology is addressing this opportunity with its on-line courses and mobile apps. The on-line courses increase students' computer literacy while using tools like Cn3D, Blast, ORF finder, and multiple databases, in directed and exploratory ways, helps students better understand biology as well gain a better appreciation for the value of the data and the field of bioinformatics. As the use of mobile devices increases, we are responding with apps like Molecule World™ that can access multiple databases and facilitate explorations into the relationships between sequence, structure, and function in fun ways.

FF0074

A comparison of hybrid assemblies: Improving Illumina-based algal draft genomes with PacBio and OpGen optical maps.

Karen Davenport, Yuliya Kunde, Olga Chertkov, and Shawn Starkenburg

Genome Science Programs, Biocience Division, Los Alamos National Laboratory

Although much effort has been expended to characterize growth phenotypes and the lipid-generation potential of many algae, knowledge of the genetic and genomic basis that defines and controls their physiological behavior is sorely lacking. Only a handful of algal genomes have been sequenced to high quality and those that have are not viable candidates for employment in biofuel production systems. Therefore, to effectively inform and populate genetic engineering pipelines, a genomic sequencing and analysis project was initiated for *Chlorella sorokiniana* 1230, a promising biofuel production strain. Assembly of Illumina paired-end reads resulted in a draft genome size of 56 Mbp. The draft genome contained 7477 contigs with an N50 of 15,504 bp. To improve the quality of the assembly, sequence reads from PacBio and optical maps from OpGen were generated. To determine which platform was more effective at improving the draft assembly, the PacBio data and the OpGen maps were assembled independently with the Illumina paired end data. Herein, we will present the comparisons of these assembly methods as well as cost-benefit analysis of each platform.

FF0076

A high-resolution, routine diagnostic, genotyping method for *Mycobacterium bovis* and *Brucella abortus* using single nucleotide polymorphisms from whole genome sequencing data.

Tod P. Stuber (Tod.P.Stuber@aphis.usda.gov)[a], Suelee Robbe-Austerman[a], Tyler Thacker[b], Patrick M. Camp[a], David T. Farrell[a], Christine R. Quance[a], Matthew M. Erdman[a]

[a]United States Department of Agriculture, Animal & Plant Health Services, Diagnostic Bacteriology Laboratory, Ames, Iowa, [b]United States Department of Agriculture, National Animal Disease Center, Ames, Iowa

A robust genotyping method is needed to provide low and high-resolution comparisons for diseases of high consequence. Here we describe a method for diagnostic laboratory use of WGS SNP data to identify the source of disease outbreaks. We describe the method used to sequence, pipeline used for data analysis, SNP verification, and interpretation and reporting. SNP comparisons are reported with an accurate, transparent and in an easy to understand format. SNP analysis from *Mycobacterium bovis* and *Brucella abortus* shows improved resolution when compared to past spoligotyping, MIRU-VNTR-24 and MLVA methods. We show SNPs discriminating herd-to-herd transmission and we discuss how unique SNPs seen in single isolates are interpreted differently than SNPs seen in multiple isolates.

DNA library preparations were done using Illumina's NexteraXT kit with 1ng of input DNA. Routine diagnostic preparations were indexed and multiple samples were ran on a single MiSeq chip using 2 X 250 paired-end chemistry. Using BWA, Picard and Broad Institute's GATK programs sequencing fragments were aligned to a reference and SNPs relative to the reference for each isolate were called. SNP calls were then placed through a pipeline using built-in Unix programs to output easy to interpret SNP comparisons of isolates as both phylogenetic trees and SNP tables. Comparisons demonstrated transmission direction from outbreak-to-outbreak and even herd-to-herd in diseases of high consequence.

This method has been effective for providing low and high resolution genotyping results using a workflow that can be done in a diagnostic laboratory setting giving comparisons in real-time to epidemiologist and field staff in an easy to interpret comparison.

Keywords: genotyping, WGS SNP analysis

FF0080

De Novo Assembly of Crithidia fasciculata Using PacBio Long Reads

Chad M. Tomlinson[1], Vincent Magrini[1], Sean McGrath[1], Amy Ly[1], Patrick Minx[1], Stephen M. Beverley[2], Wesley C. Warren[1], and Richard K. Wilson[1]

[1] The Genome Institute at Washington University School of Medicine [2] Department of Molecular Microbiology, Washington University School of Medicine

Our group is interested in utilizing PacBio long reads to improve upon existing reference assemblies produced primarily using short reads. The current GenBank reference for Crithidia fasciculata is an example of a fragmented assembly from short read data. This assembly is 40.2 Mb in size and is in 4,377 scaffolds and 5,535 contigs, with a scaffold N50 length of 113,621 bp and a contig N50 length of 42,099 bp. In an effort to improve upon this assembly, we generated four SMRT cells of P4/C3 10 Kb data and eight SMRT cells of P5/C4 20 Kb data. One challenge with de novo assembly of PacBio long reads is the need to reduce the approximate 15% randomly distributed error observed across single-pass reads. We tested the HGAP and ECtools algorithms for error-correction and found that ECtools yielded a larger amount of error-corrected data and a more complete and contiguous final assembly than error-correction using HGAP. The largest yield of error-corrected coverage that we were able to obtain from HGAP was only 17.79X. Error-correction of the data using ECtools resulted in a more impressive 49.4X yield of error-corrected data. ECtools uses high identity sequence such as unitigs or contigs from a reference assembly to error-correct the PacBio long reads. We used the contigs from the current GenBank Crithidia reference in conjunction with ECtools to error-correct the Crithidia data. We obtained the most optimal assembly by extracting 30X of the longest reads from this error-corrected dataset and assembling the data using Celera 8.1. This resulted in a 40.3 Mb assembly in 589 contigs with a contig N50 length of 308,523 bp. This represents a nearly ten-fold reduction in the number of contigs and a nearly eight-fold improvement in the contig N50 length compared with the current GenBank reference.

Keywords: Genome Assembly, Genome Improvement, PacBio

FF0084

Evaluation and validation of de novo and hybrid assemblies to derive high quality genome sequences

S. M. Utturkar[1], D. M. Klingeman[2], W. Mitchell[3], M. L. Land[2], S. De Tissera[3], S. Segovia[3], M. Köpke[3], C. W. Schadt[2], M. J. Doktycz[2], D. A. Pelletier[2], S. D. Brown[2]

[1]University of Tennessee, Knoxville, TN, [2]Oak Ridge National Lab., Oak Ridge, TN, 3LanzaTech NZ, Ltd, Auckland, NEW ZEALAND.

Background: Multiple copies of ribosomal RNA (rRNA) operons (containing 16S, 23S and 5S sequences) present one of the greatest technical challenges during the bacterial genome assembly process. The aim of the present study was to assess the potential of next generation sequencing (NGS) technologies, different *de novo* and hybrid assembly approaches to resolve rRNA operons and obtain the most accurate assemblies.
Methods: Sequencing of 4 novel bacteria was performed using Illumina paired-end (PE), mate-pair (MP), Roche 454 and PacBio RS-I technologies. Assemblies were performed using combinations of NGS platforms with various de novo and hybrid assembly protocols. Quality assessments were performed through *in silico* evaluation tools. In absence of finished reference sequence, the *in silico* evaluations were only useful to rank the assemblies without true knowledge of their accuracy. Predictions of different rRNA operon contents were tested and validated by PCR and Sanger sequencing.

Results: Assembly of Illumina PE reads only generated partial rRNA operons. The hybrid assembly of PE-MP or PE-454 reads led to predictions of complete rRNA operons, but their 3' and 5' flanking regions were missing. Incorporation of long PacBio RS-I reads extended the regions flanking rRNA operons, which enabled predictions of multiple copies of rRNA operons, their genomic positions and overall assembly quality was improved up to 'non-contiguous finished' status. The rRNA analysis provided additional confidence for assembly accuracy by validating multiple copies of rRNA operons. In contrast to these hybrid approaches, we were able to obtain the complete genome sequence of class III complexity bacterium (*Clostridium autoethanogenum*) using only PacBio RS-II technology and without need of manual finishing. The comparison of the finished genome against draft and hybrid assembly revealed the limitation of Illumina/454 technologies to resolve nine copies of rRNA operons.
Conclusion: rRNA operon predictions provide an additional criterion for assembly quality matrices.

Acknowledgement: Research sponsored in part by the Genomic Science Program, U.S. Department of Energy, Office of Science, Biological and Environmental Research and by LanzaTech NZ Ltd.

FF0086

Rapid cloud-based data processing and analysis of >15,000 whole exomes in a collaborative setting promotes novel gene discovery

Narayanan Veeraraghavan[1], Andrew Carroll[2], Shalini Jhangiani[1], Alexander Li[4], Tomasz Gambin[3], Zhuoyi Huang[1], Ginger Metcalf[1], Fuli Yu[1], Alanna Morrison[4], Donna Muzny[1], Richard Daly[2], James Lupski[3], Geoff Duyk[2], Richard Gibbs[1,3], Eric Boerwinkle[1,4]

[1]Human Genome Sequencing Center, Baylor College of Medicine, [2]DNAnexus, [3]Department of Molecular and Human Genetics, Baylor College of Medicine, [4]Human Genetics Center, University of Texas Health Science Center at Houston

We are now witnessing NGS-based efforts at consortium scales involving hundreds of collaborators and thousands of study subjects, to exhaustively characterize diseases and involved genes. There are three critical areas that serve as enablers for efficient genomic discovery and diagnostics: (a) large scale computational resources, (b) rapid and flexible analysis capabilities, and (c) efficient data delivery and consumption. To address these, we have entered into collaboration between DNAnexus, the Baylor-Hopkins Center for Mendelian Genomics (CMG), and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium, to create a Gene Discovery Commons for both Mendelian (CMG) and common chronic (CHARGE) diseases.

As a test pilot of the Commons, we analyzed 1000 cases from the CMG spanning 15 diseases, and 3500 whole genome and 11000 whole exome sequences of deeply phenotyped individuals from the CHARGE consortium. The CHARGE cohort samples provide a large comparison group for the CMG, thus increasing statistical power.

Rapid deployment of our Mercury NGS pipeline to Amazon cloud (AWS) was made possible through the DNAnexus platform. Some of the key features of the computational framework are: zero set-up, on-demand scale-up, version control, reproducibility, visualization, and compatibility with off-the-shelf third party tools, naturally enforced standards, compliance with CAP/CLIA/HIPAA, and ISO27001 data handling and security compliance. The ability to easily, quickly and securely share data between collaborators is yet another hallmark of this framework.

As an example of scientific utility, we sequenced a sample of 204 probands with heterotaxy, a syndrome involving malarrangement of internal organs within the chest and abdomen. Analysis of these data identified 38 genes including known heterotaxy genes and novel candidates.

These and other examples demonstrate the utility of the Gene Discovery Commons, and the effectiveness of a cloud base framework, for promoting collaboration, corroboration and clinical translation.

Keywords: cloud computing, data management, bioinformatics framework, gene discovery, collaborative science. *Also to be presented as talk.*

FF0087

Transcriptome Sequencing of Forensically Relevant Fluids and Tissues to Optimize Degradation Analysis for Sample Age Estimation

Kate Weinbrecht, MS and Robert Allen, PhD

Oklahoma State University Center for Health Sciences; School of Forensic Sciences

Research on the forensic applications of RNA analysis has increased greatly in the last decade. The use of RNA in forensic analysis offers investigators a broader amount of information from a biological sample than can be obtained from DNA analysis alone. Applications include the use of RNA to identify tissue identity, age, and disease status. Although recent research has indicated many possible forensic applications of RNA analysis, many questions remain concerning the behavior of RNA in degraded and limitedly available samples. Specifically, there remains to be a thorough understanding of the mechanism of RNA degradation in post-mortem or deposited samples. Thus, choosing suitable RNA markers for evaluating the approximate age and identity of forensic samples can be problematic. The purpose of this research is to evaluate RNA degradation in forensically relevant biological fluid types (blood, saliva, semen, and vaginal fluid) in order to establish tissue-specific transcriptome degradation profiles and patterns that may be used to establish approximate sample age. Whole transcriptome sequencing of RNA isolated from samples that have been aged up to one year was performed to evaluate the patterns of RNA degradation in relation to sample age. Initial results indicate both whole transcriptome degradation and individual transcript degradation profiles that correlate with sample age in each of the body fluid types examined.

Keywords: RNA-seq, Degradation, Forensic, Age, Transcriptome

FF0088

µFORGE: A Consortium Dedicated to Advancing Microbial Forensics through Genomics.

Richard Winegar

MRIGlobal

Microbial forensics is an emerging discipline that applies science to the investigation of biocrimes. A primary objective is attributing biological evidence to a particular group or individual through investigative leads generated by comparing genetic information of isolates used in a crime with isolates from potential sources. However, the ability to draw conclusions about the relatedness between isolates may be limited when using traditional genotyping methods due to insufficient resolution. Whole genome sequencing provides ultimate resolution, but requires comparative databases against which to compare an isolate in the context of an organism's true genetic diversity.

DHS has funded a program to improve existing genome databases for microbial forensics use. In response, we have assembled the Consortium for Microbial Forensics and Genomics (µFORGE). Members include:

- MRIGlobal, a leading non-profit research institute with expertise in biodefense and microbial forensics.
- UTMB, a premier academic research organization with preeminent collections and expertise in biothreat viruses and maximum containment facilities
- OSU, whose National Institute for Microbial Forensics & Food and Agricultural Biosecurity is the only US research entity focused on plant pathogen forensics
- cBio, a bioinformatics company with extensive experience in genome assembly and finishing as well as building and maintaining databases

DHS and other government stakeholders have identified over 70 pathogens that pose a significant threat to humans or agriculture. In Phase I, the µFORGE team will 1) survey available genome sequences for each organism and identify gaps in genetic diversity; 2) prioritize organisms for sequencing and surveillance; 3) work with DHS, NCBI and other stakeholders to establish database requirements; and 4) initiate genome sequencing. Phase II is anticipated to be a continuing effort to identify, acquire, and sequence genomes of high priority pathogens. High-quality genome sequences will be submitted to NCBI to provide a well-curated public resource to support the microbial forensics community.

Keywords: Forensics, BioSurveillance, Select Agents, Next Generation Sequencing, bioterrorism

FF0090

GRCh38: Updating the Human Reference Genome assembly.

Jonathan Wood and Kerstin Howe on behalf of the Genome Reference Consortium.

The Reference Human Genome is an indispensable tool for clinical and fundamental research and its accuracy is of prime importance.

Since the GRCh37 release in 2009, 170 assembly patches affecting 3% of the chromosome sequence have been added to the genome, comprising more than 6 Mb of novel sequence.

There were however, several challenges remaining; unplaced and missing sequence, misassembled highly repetitive regions, the absence of centromere data, along with the presence of large numbers of erroneous bases which cause problems for NGS short read alignment and incorrect gene annotation.

To address these issues and further improve the accuracy of the reference genome we undertook a renewed effort to integrate analysis from different sources. We will present our progress on how performing our own analyses plus utilising analyses from collaborators and publicly available sequence data led to corrections in the reference genome. Optical Mapping, Strand Sequencing and Admixture Mapping in addition to the retiling of complex regions in a single haplotype have greatly contributed to improvements in the chromosomal structure and have allowed confident sequence placement. Large-scale correction of erroneous bases using 1000 Genomes data has improved gene annotation genome-wide. Furthermore we added modelled centromeres, which along with the inclusion of WGS assemblies greatly improve the alignment of NGS short read sequences without the need for decoy sets.

All of this new data together with all previous patches are now integrated into the latest major assembly update of the reference human genome, GRCh38, released in December 2013. As a result of this combined effort there is a significant improvement in both the structure and the representation of variation in the chromosomes.

We believe the new GRCh38 assembly will greatly benefit both researchers and the clinical community and that the reference human genome is one step closer to completion.

FF0093

Genetic Characterization of Extended-Spectrum β-Lactamase Genes of *Escherichia coli* in Georgia

T. Tevdoradze[1], E. Zhgenti[1], D. Zorikov[1], G. Chanturia[1], Matthew Scholz[2], Patrick Chain[2], Tracy Erkkila[2]

[1] National Center for Disease Control and Public Health (NCDC), Tbilisi, Georgia [2] Los Alamos National Laboratory, NM, USA

Escherichia coli is the species most frequently associated with clinical infections by extended- spectrum-β-lactamase (ESBL) producing isolates. ESBL *E. coli* as a multidrug resistant pathogen represents a major problem in human and veterinary medicine. CTX-M enzymes are a group of class A extended-spectrum β-lactamases (ESBLs) that become the most prevalent ESBL enzymes in clinical *Enterobacteriaceae* isolates, especially in ESBL-producing *Escherichia coli* in Europe, Asia, and South America. More than 130 CTX-M allelic variants have been described, divided into five groups depending on their amino acid sequence. *E. coli* isolates with particular CTX-M subtypes associated with different geographic regions. There is generally a lack of comprehensive data regarding the prevalence and genetic characteristics of ESBL-producing *E. coli* in Georgia.

In this study we characterized two ESBL *E. coli* strains isolated from the post-surgical samples obtained from gallbladder of the patients with chronic calculous cholecystitis. Pathogens were isolated on selective and differential media followed by identification by API-20E system. Sequence of isolates were performed by the next-generation sequencing (NGS) Illumina MiSeq Platform. Sequencing reads were further analyzed by the CLC Genomics Workbench software package (CLC bio).

In this study, we determined the CTX-M subtype of isolates by nucleotide sequencing. Sequence types also were identified based on standard 7 housekeeping loci (adk, fumC, gyrB, icd, mdh, purA, and recA), selected from the *E. coli* database developed by M. Achtamn et al (http://mlst.ucc.ie/). The allele and profile assignments were determined on the basis of the central *E. coli* MLST database at the above website. The present study is the first attempt to characterize the resistance genes and MLST characteristics of ESBL-producing *E. coli* strains in Georgia.

Keywords: ESBL-Escherichia coli, CTX-M, Sequence, MLST

FF0094

Data Release for Polymorphic Genome Assembly Algorithm Development

Marty Badgett (mbadgett@pacificbiosciences.com)

Pacific Biosciences

Heterozygous and highly polymorphic diploid (2n) and higher polyploidy (n > 2) genomes have proven to be very difficult to assemble. One key to the successful assembly and phasing of polymorphic genomics is the very long read length (9-40kb) provided by the Pacific Biosciences® RS II system. We recently released software and methods that facilitate the assembly and phasing of genomes with ploidy levels equal to or greater than 2n (presentation abstract by Jason Chin). In an effort to collaborate and spur on algorithm development for assembly and phasing of heterozygous polymorphic genomes we have recently released sequencing datasets that can be used to test and develop highly polymorphic diploid and polyploidy assembly and phasing algorithms. These data sets include multiple species and ecotypes of Arabidopsis that can be combined to create synthetic in-silico F1 hybrids with varying levels of heterozygosity. Because the sequence of each individual line was generated independently, the data set provides a 'ground truth' answer for the expected results allowing the evaluation of assembly algorithms. The sequencing data, assembly of inbred and in-silico heterozygous samples (n=>2) and phasing statistics will be presented. The raw and processed data has been made available to aid other groups in the development of phasing and assembly algorithms.

Keywords: Heterozygous diploid Assembly Data Release

FF0098

Workflows Incorporating a DNA enzyme repair mix improve NGS library prep from FFPE samples

Fiona J Stewart, Lixin Chen, Laurence Ettwiller, Pingfang Liu, Eileen T Dimalanta, and Thomas C Evans Jr.

Treating biopsy samples with formalin and embedding them in paraffin is a widely practiced method for preserving and archiving clinical samples. The rise of next generation sequencing (NGS) technologies makes it possible for the billions of unique formalin-fixed, paraffin-embedded (FFPE) samples stored worldwide to provide a wealth of information in retrospective genomic studies of human disease. Although well suited to histopathological studies, it has been challenging to retrieve genetic information from FFPE samples. This is often attributed to DNA damage incurred during fixation, including fragmentation, oxidation, deamination, and protein-DNA crosslinks. The poor quality of DNA extracted from FFPE samples has significantly limited the information that can be generated by NGS technologies.

The PreCR Repair Mix is an enzyme cocktail formulated to repair damaged template DNA prior to its use in subsequent detection technologies. PreCR is active on a broad range of DNA damages, including modified bases, nicks and gaps, and a variety of blocking moieties at the 3´end of DNA. In this study, we have investigated the effects of PreCR-mediated DNA repair on NGS library preparation from FFPE samples, using a longer library preparation method that has separate PreCR, end repair, dA tailing and adaptor ligation steps, as well as more streamlined protocol that combines reactions and reduces bead cleanup steps. The PreCR pre-treatment of DNA samples was found to increase library yield and library success rates without introducing bias into the sequence data. In conclusion, incorporating PreCR treatment into library preparation workflows improves the quantity and quality of NGS libraries from FFPE samples. Furthermore, streamlined protocols that combine reaction steps significantly reduce the turn-around time enabling high throughput processing of samples for clinical analysis and large scale genomic studies.

FF0100

Host-Pathogen Interaction Dynamics of Human Astrocytes Infected with Venezuelan Equine Encephalitis Virus

Alan Baer[1], Danielle Swales[2], Nicole Waybright[2], Lindsay Lundberg[1], Jonathan Dinman[3], Jonathan Jacobs*[2], Kylene Kehn-Hall*[1]

[1]National Center for Biodefense and Infectious Diseases, George Mason University, Manassas, Virginia [2]MRIGlobal, Biosurveillance and Global Health Security, Rockville, Maryland
[3]University of Maryland, Department of Cell Biology and Molecular Genetics, College Park, Maryland

Venezuelan Equine Encephalitis Virus (VEEV) is an emerging arthropod-borne virus responsible for causing acute encephalitis, and often death, in animal and human hosts. VEEV is a priority pathogen that has previously been weaponized. Due to its continued environmental persistence in the Americas, it represents a significant threat to U.S. public health and economic security. The increased circulation and spread in the Americas of VEEV, and other encephalitic arboviruses such as Eastern Equine Encephalitis Virus (EEEV) and West Nile virus (WNV), underscores the need for research aimed at characterizing the pathogenesis of viral encephalomyelitis as a foundation for the development of novel medical countermeasures. In this study, we have sought to characterize the host-pathogen dynamics of VEEV in the human neuronal cell line U87MG by carrying out RNA sequencing of poly(A)+ mRNAs. We aim to identify critical alterations in the host transcriptome that take place within the first 24 hours following VEEV infection. Triplicate samples were collected at 4, 8, and 16 hours post-infection and RNA-Seq data acquired using an Ion Torrent PGM. Significant differentially expressed genes were part of the following super pathways: immune response IFN alpha/beta signaling, immune response IL-2 activation and signaling, regulation of nuclear SMAD2/3 signaling, and development glucocorticoid receptor signaling. Specifically we observed an increase in interferon regulated genes IFIT1, IFIT2, IFIT3, and OASL following VEEV infection. We also observed an increase in EGR-1 and differential expression of a number of genes that are involved in the EGR-1 pathway including ADAMT21, ATF3, KLF6, MYC, JUN, and PTGS2. Data from these studies will be leveraged towards identifying specific host mRNA transcripts or pathways suitable for therapeutic intervention, as well as provide mechanistic details regarding how alphaviruses manipulate the host transcriptome to facilitate replication.

FF0102

Hepatitis B Virus Recombinants Circulating among the Acute and Chronic Viral Hepatitis patients in Kenya

Ochwoto M (omissiani@gmail.com)[1,4], Songok E.[1,5], Kimotho, J.H.[1], Okoth F.A.[1], Oyugi J.[2,5], Kiptoo M.[1,3], Ng'ang'a Z.[3], Budambula N.[3], Giles E.[4], Andonov A[4,5], Osiowy C.,[4,5]

[1]Kenya Medical Research Institute (KEMRI); [2]University of Nairobi Medical Microbiology Department, [3]Jomo Kenyatta University of Agriculture and Technology (ITROMID), [4] Public Health Agency of Canada, National Microbiology Laboratory (NML), [5]University of Manitoba (UofM)

There are five types of viral hepatitis; A-E that differ in distribution, transmission, disease burden, manifestation and progression. They account for the majority of liver related morbidity and mortality worldwide. In order to determine the prevalence and molecular characterization of these hepatitis viruses (A-E) among patients with jaundice seeking medical services in Kenya, 395 patients were recruited in four selected hospitals. Approximately 7ml of blood was drawn from the patients for serology of Hepatitis A Virus (HAV), Hepatitis C Virus (HCV), Hepatitis D Virus (HDV), and Hepatitis E virus (HEV) IgG/IgM antibodies, and Hepatitis B surface and core antigens (HBsAg and HBcAg). Nucleic acid from all positive samples was extracted, amplified and sequenced.

Acute HAV infection was detected in 6.1% among 395 patients. The prevalence was high among two cities; Nairobi (54.2%) and Kisumu (29.2%). Sequences of HAV revealed that there were two different circulating strains of genotype 1b in 2010-2013. No serological evidence of HDV and HEV infection was observed. Screening for HCV infection by ELISA identified 3.8% of samples to be weak positive; however none were confirmed using the immunoblot HCV Inno-Lia assay. HBV was the most prevalent Hepatitis virus, with 29.87% HBsAg positivity. HBV genotype A1 was predominant (89.6%) followed by D (10.4%). Full genome analysis of HBV/D isolates showed that a third of them belonged to a putative new circulating sub-genotype having > 4% nucleotide divergence from both subgenotypes D7 and D8. Another third were D8 (D/E recombinant) and the rest were genotype D7. All Genotype HBV/D were located in the Western Kenya region among male patients 40 years of age and above. Within the antigenic determinant region (aa100-160), no mutations were found with HBV/D sequences whereas in HBV/A1; sG112R, sT114S sF134V, sT143M, sM103I, sQ129R, sG130N, and sP135H mutations were observed. BCP/Pre-core mutations A1762T/G1764A were common with HBV/A1 (32.43%) and none with HBV/D; however, the stop codon mutation ntG1896A and C/T1858 were 3.90% among HBV/A1. In conclusion, genotype 1b could be responsible for HAV outbreaks in 2010-2013. The high prevalence of HBV and HAV, despite the availability of a stable vaccine, may have been as a result of population selection. There is need to document the clinical relevance, manifestation and distribution of the unique HBV/D virus. The mutations within HBV aa100-160 might interfere with diagnosis whereas BCP/Pre-core mutations A1762T/G1764A have been constantly associated with Hepatocellular carcinoma.

Keywords: Hepatitis B Genotypes, Hepatitis A virus, BCP/Pre-core mutations, Antigenic Determinant mutations

FF0103

Next Generation Sequencing of HIV strains from a drug naïve population in North Rift Kenya show a high prevalence of low abundant drug resistant variants.

Songok EM, Cheriro W, Brooks J, Liang B, Hezhao J, Lihana R, Kiptoo M

The advent of antiretroviral treatment (ART) has resulted in dramatic reduction in AIDS related morbidity and mortality. However the emergence and spread of antiretroviral drug resistance (DR) threatens to negatively impact on treatment regimens and compromise efforts to control the epidemic. It is recommended that surveillance of drug resistance occur in conjunction with scale-up efforts to ensure appropriate first-line therapy is offered relative to the resistance that exists. However standard resistance testing methods used in Subsahara Africa rely on techniques that miss out on low abundance DR variants (LADRVs) which have been documented to contribute to treatment failure. The use of next generation sequencing (NGS) has been shown to be more sensitive for LADRVS. We have carried out a preliminary investigation using NGS to determine the prevalence of LDRVS among a drug naïve population in North Rift Kenya.

Antiretroviral naïve patients attending a care clinic at Moi Teaching and Referral Hospital (MTRH) in Eldoret, Kenya were requested and with consent provided blood samples for DR analysis. DNA was extracted, amplified and nested PCR conducted on pol RT region using with primers tagged with multiplex identifiers (MID). Resulting PCR amplicons were purified, quantified and pyrosequenced using a GS FLX Titanium PicoTiterPlate (Roche). Valid pyrosequencing reads were aligned with HXB-2 and the frequency and distribution of nucleotide and amino acid changes determined using an in-house Perl script. DR mutations were identified using the IAS-USA HIV DR mutation database.

Sixty samples were successfully sequence of which 25 were subtype A, 11 subtype D, 1 Subtype C and the remaining were recombinants. Forty six (76.6%) had at least one drug resistance mutation; with 25 (41.6%) indicated as major and the rest 21 (35%) indicated as minor. The most prevalent mutation was NRTI position K219Q/R (11 of 46, 24%) followed by NRTI M184V (5 of 46, 11%) and NNRTI K103N (4 of 46,9%).
Our use of NGS technology revealed a high prevalence of LADRVs among drug naive populations in Kenya. The impact of these mutations on clinical outcome on ART can only be ascertain through a long term follow-up.

FF0104

Research Using Rapid Assessment of AML Genes for Mutation Detection

Cristina Van Loy, Mark Andersen, Kate Rhodes, Steve Roman, Adam Broomer, Michael Allen, Denise Topacio, Guoying Liu, Manimozhi Manivannan, Charles Scafe, Fiona Hyland, Chrysanthi Ainali, Alexander Sartori, AML Core Network, Daniel Mazur, Anelia Kraltcheva, Eileen Tozer, Guobin Luo, Mindy Landes, Sihong Chen, Josh Shirley, Kevin Heinemann.

Thermo Fisher Scientific, Carlsbad, CA

Targeted sequencing using the Ion AmpliSeq™ Library kit combined with the Ion PGM™ sequencing instrument is a fast and effective research method to identify genetic variants in tumor samples. A new targeted primer panel for amplification of genes involved in Acute Myeloid Leukemia (AML) has been developed by Life Technologies. The panel covers 19 genes characterized using 237 specific primer pairs in two highly multiplexed PCRs. To demonstrate the coverage efficiency, we evaluated libraries prepared from isolated genomic DNA. When libraries from 4 samples were run on a single Ion 318™ Chip, the average coverage depth was >3000x, with >97% of the target bases covered >500X. Additionally, >80% of reads were on-target. The panel was tested on control samples and analyzed using the Ion Reporter™ Software, and expected variants were detected with high sensitivity and specificity. In addition, improvements to template preparation and sequencing with the Ion Hi-Q™ Sequencing Chemistry (unreleased product in development) resulted in a higher percentage of end to end reads and uniformity, as well as reduced strand bias.

FF0105

Using Comparative Genomic Analyses to Determine the Pan-Genome and Genetic Structure of *Staphylococcus aureus*.

CC Roe, D Lemmer, M Valentine, E Driebe, PS Keim, DE Engelthaler

Translational Genomics Research Institute, Flagstaff, AZ

Next generation sequencing has allowed for the rapid genomic description of some bacterial species. Analysis of a pan-genome provides the breadth of diversity present in a single species and when studying pathogens, can further elucidate genetic markers for disease. For this study, we analyzed 250 diverse *Staphylococcus aureus* genomes to better understand the genetic landscape, structure and diversity of the organism at a strain level. Evidence of homologous recombination throughout the *S. aureus* genome was observed using an in-house, phylogenetic SNP analysis pipeline (NASP) and measuring the phi statistic as well as visualization in Splitstree. Draft genomes were produced using a modified SPAdes assembler and the pan-genome was constructed using Large Scale Blast Score Ratio (LS-BSR), which compares predicted gene content across the entire sample set. We identified the core *S. aureus* genome as having 1,642 genes while the variable genome is composed of 4,821 genes. We also identified gene content specific to each *S. aureus* lineage. A large proportion of genes associated with pathogenicity were not shared by all *S. aureus* strains, suggesting this is an inadequate measurement for determining virulence in the species as a whole but could account for differences in virulence across lineages. *S. aureus* has an open pan-genome with substantial diversity within as well as between lineages. The scope of genetic variability across *S. aureus* provides insight into the species' ability to thrive in a diverse range of environmental niches.

Keywords: Pan-Genome, BSR, *Staphylococcus aureus*, whole genome sequencing, SPAdes assembler

FF0106

Hotspot mutation and fusion transcript detection from the same nonsmall lung adenocarcinoma sample

Angie Cheng, Varun Bagai, Joey Cienfuegos, Natalie Hernandez, Mu Li, Jeff Schageman, Richard Fekete, Rosella Petraroli, Alexander Vlassov, and Susan Magdaleno

Life Technologies

The presence of certain chromosomal rearrangements and the subsequent fusion gene derived from translocations has been implicated in a number of cancers. Hundreds of translocations have been described in the literature recently but the need to efficiently detect and further characterize these chromosomal translocations is growing exponentially. The two main methods to identify and monitor translocations, fluorescent in situ hybridization (FISH) and comparative genomic hybridization (CGH) are challenging, labor intensive, the information obtained is limited, and sensitivity is rather low. Common sample types for these analyses are biopsies or small tumors, which are very limited in material making the downstream measurement of more than one analyte rather difficult; obtaining another biopsy, using a different section or splitting the sample can raise issues of tumor heterogeneity. The ability to study mutation status as well as measuring fusion transcript expression from the same sample is powerful because you’re maximizing the information obtained from a single precious sample and eliminating any sample to sample variation. Here we describe the efficient isolation of two valuable analytes, RNA and DNA, from the same starting sample without splitting, followed by versatile and informative downstream analysis. This methodology has been applied to FFPE and degraded samples as well as fresh tissues, cells and blood. DNA and RNA were recovered from the same non-small lung adenocarcinoma sample and both mutation analysis, as well as fusion transcript detection was performed using the Ion Torrent PGM™ platform on the same Ion 318™ chip. Using 10ng of DNA and 10ng of RNA input, we applied the Ion AmpliSeq™ Colon and Lung Cancer panel to analyze over 500 COSMIC mutations in 22 genes and the Ion AmpliSeq™ RNA Lung Fusion panel to detect 40 different fusion transcripts.

FF0107

Molecular epidemiology of C*ryptosporidium* and G*iardia* species around humans, livestock and wildlife in Kibale National Park

[1]Arinaitwe Eugene (arieugene@yahoo.com), Innocent Rwego, Benon Asiimwe, Joloba Moses, [1]Rose Ademun, [1]Deo Ndumu and [1]Chrisostom Ayebazibwe.

[1]National Animal Disease Diagnostics and Epidemiology Centre, Ministry of Agriculture Animal Industry and Fisheries, P.O.Box 513 Entebbe. [2]Makerere University, P.O.Box 706 Kampala, Uganda.

Cryptosporidiosis and giardiasis are among the most economically important Zoonotic diseases worldwide. This study was carried out around Kibale National Park in a village called Kanyawara, to determine the prevalence of *Cryptosporidium* and *Giardia* species and the circulating serotypes in humans, livestock and baboons. PCR-RFLP assay was used to establish the genotypes. Cryptosporidium Oocyst Wall Protein (COWP) and Small Subunit Ribosomal RNA (SSUrRNA) genes were used to detect species-specific sequences for *Cryptosporidium spp.* Beta giardin and TPI genes were used to detect species- specific sequences for *Giardia*. The *Cryptosporidium spp* found in both humans and baboons were identical and the one isolated from a goat was unique, by comparing their sequences. The prevalence of *Cryptosporidium species* was, 3.6 % in Humans, 2.6 % in Livestock and 1.9 % in Baboons while that of *Giardia spp.* was 3.6 %, 5.1 % and 0 % in humans, livestock and baboons respectively. The parasites were found in all age groups of humans. Co-infection with *Cryptosporidium* and *Giardia* was detected in a goat. By restriction digestion and sequencing analysis of COWP gene of *Cryptosporidium* showed that the parasites were 100 % related to *Cryptosporidium parvum.* Restriction digestion and sequencing analysis of beta giardin and TPI genes of *giardia* show that the parasites were *Giardia Intestinalis.* By this study we have shown that *Cryptosporidium parvum and Giardia Intestinalis* (assemblages B and E) are parasites circulating in Kanyawara village. Co-infection with livestock and wildlife posses a continuous risk of infection to humans and control measures should take this into consideration during community health activities in the area.

Keywords: Molecular epidemiology, *Cryptosporidium, Giardia*, sequencing, Restriction digestion.

FF0109

Comparison of six *Francisella tularensis* draft genomes from country of Georgia

G. Chanturia[1], D. Zorikov[1], M. Scholz[2,3], P. Chain[2], T. Erkilla[2]

1 National Center for Disease Control and Public Health of Georgia, Richard G. Lugar Center for Public Health Research, Georgia; 2 Los Alamos National Laboratory, NM, USA; 3 Michigan State University, MI, USA

Francisella tularensis is select agent that is spread in two foci at the eastern part of the country of Georgia. The active surveillance on this pathogen is carried out by the National Center for Disease Control and Public Health of Georgia (NCDC) - the significant strain collection existed recently at NCDC - Richard G.Lugar Center for Public Health Research is the result of the continuous seasonal field sampling. The ongoing DTRA project "Epidemiology and Ecology of Tularemia in Georgia" goals deeper investigation of tularemia ecological niche at known foci and other regions of the country.

The SNP and MLVA genotyping of archival strains was already performed in collaboration with Northern Arizona University (NAU) and Walter Reed Army Institute of Research (WRAIR). SNP typing was based on whole genome sequencing of one strain from NCDC collection and comparison to other *F. tularensis* strain genomes that revealed Georgian strain as a basal for the most of European clades. Discovery of country specific SNPs and use for typing and phylogenetic analysis of the rest of the strains accomplished the study.

Six *F. tularensis* strains from NCDC collection selected from different SNP and MLVA profiles were sequenced using Illumina MiSeq – the Next Generation Sequencing platform at the NCDC-Lugar Center. Several bioinformatics tools were used for data processing and phylogenetic analysis. *F. tularensis* Live Vaccine Strain (LVS) genome was used for alignment. Analysis revealed the new set of SNPs and deletions among Georgian *F. tularensis* strains. Most of them belonged to genetically distant strains represented the basal and the last subclades of Georgian SNP lineage.

The results obtained from sequencing will be used for SNP discovery and genetic characterization of the new strains isolated in different regions of Georgia and will help to improve diagnostic and genotyping capabilities at NCDC–Lugar Center.

Keywords: *F. tularensis*, Georgia, NGS, Phylogeny

FF0115

Molecular Indexing for Improved RNA-Seq and ChIP-Seq

Dr Masoud Toloue

Bioo Scientific

Most current Next Generation Sequencing (NGS) library prep methods introduce significant sequence bias. The use of enzyme processing and fragmentation steps can introduce errors in the form of incorrect sequence and misrepresented copy number. Conventional library construction involves the ligation of a population of cDNA or ChIP DNA molecules with adapters prior to amplification and sequencing. An inherent weakness of conventional RNA-Seq and ChIP-Seq analysis is that molecules that amplify more efficiently will unavoidably result in a higher number of reads than molecules that do not amplify as well during the library construction PCR step. Therefore, when multiple reads mapping to the same molecule are encountered, it is not possible to determine whether sequenced reads originate from the same or different molecules. With molecular indexed libraries, each molecule is tagged with a molecular index randomly chosen from ~10,000 combinations so that any two identical molecules become distinguishable (with odds of 10,000/1), and can be independently evaluated in later data analysis. Analysis using molecular indexing information provides an absolute, digital measurement of gene expression levels, irrespective of common amplification distortions observed in many RNA-Seq and ChIP-Seq experiments. This type of indexing requires no additional steps in workflow and increases the precision of downstream analysis. At low sequencing depths, analysis with use of molecular indices is identical to conventional analysis and generates equivalent RPKM values in all applications. As sequencing depth increases, individual molecular resolution also increases. In quantitative RNA-Seq and ChIP-Seq experiments, the molecular indices distinguish re-sampling of the same molecule from sampling of a different molecule. At high sequencing depths, each molecule can be distinguished and the entire library can be analyzed to provide absolute numbers of each molecule. Resolving individual clones of molecules is critical for increasing sequencing accuracy, measuring bias, PCR duplication rates and identifying mutations in complex sample types. While it is well known that library prep methods introduce bias, tools for measuring it are needed if we are to start using NGS for accurate and quantitative gene expression measurements. Toward achieving that goal, we propose the use of molecular indices for all RNA-Seq and ChIP-Seq experiments.

Poster Session Notes

Poster Session Notes

05/29/2014 - Thursday				
Time	Type	Abstract #	Title	Speaker
7:30 - 8:30am	Breakfast	x	La Fonda Breakfast Buffet	Sponsored by NEB
8:30 - 8:45	Intro	x	Welcome Intro from Los Alamos National Laboratory	TBD
x	Session Chair	x	Session Chairs	Chair - Johar Ali Chair - Donna Muzny
8:45 - 9:30	Keynote	FF0111	Dissecting the Missing Diagnostic Yield in exome sequencing	Deanna Church Sponsored by Personalis
9:30 - 10:00	Speaker 1	FF0057	A de novo Whole Genome Shotgun Assembler for Long Reads	Gene Myers
10:00 - 10:15	Speaker 2	FF0044	Speeding up NGS software development	D. Lavenier
10:15 - 10:30	Speaker 3	FF0038	New Frontiers of Genome Assembly with SPAdes 3.1	Anton Korobeynikov
10:30 - 11:00am	Break	x	Beverages and Snacks Provided	Sponsored by CLC Bio
11:00 - 11:15	Speaker 4	FF0065	Upgrading large genomes using Pacific Biosciences long reads and PBJelly software	Jeffrey Rogers
11:15 - 11:30	Speaker 5	FF0014	Assembly and Phasing of Polymorphic Heterozygous Diploid Genomes	Jason Chin
11:30 - 11:45	Speaker 6	FF0063	De novo Assembly of Medicago truncatula Genome Lines Using Illumina and Pacific Biosciences Sequencing Technologies	Thiruvarangan Ramaraj
11:45 - 12:00	Speaker 7	FF0062	Illumina sequencing with no artifacts	Zbyszek Otwinowski
12:00 - 12:15	Speaker 8	FF0034	Near perfect de novo assemblies of eukaryotic genomes using PacBio long read sequencing	James Gurtowski
12:15 - 1:55pm	Lunch	x	New Mexican Lunch Buffet	Sponsored by Promega
x	Session Chair	x	Session Chairs	Chair - Tina Graves Chair - Bob Fulton
1:55 - 2:10	Speaker 9	FF0069	Anchored Assembly: An Algorithm for Large Structural Variant Detection Using NGS Data	Niranjan Shekar
2:10 - 2:20	Speaker 10	FF0085	De Novo Assembly and Structural Analysis Using Extremely Long Single-Molecule Imaging	Han Cao
2:20 - 2:30	Speaker 11	FF0020	Human sequence assembly scaffolding using Irys genome maps	Heng Dai
2:30 - 2:45	Speaker 12	FF0071	Efficient de novo assembly of long NGS reads	Martin Simonsen
2:45 - 3:00	Speaker 13	FF0047	*de novo* mammalian assembly of one-library PCR-free 250-base Illumina reads	Iain MacCallum
3:00 - 3:15	Speaker 14	FF0032	Creating a Single Haplotype Human Genome Assembly	Tina Graves-Lindsay
3:15 - 3:45pm	Break	x	Beverages and Snacks Provided	Sponsored by BioNano
x	Session Chair	x	Session Chairs	Chair - Alla Lapidus Chair - Darren Grafham
3:45 - 4:00	Speaker 15	FF0033	Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)	Stephanie Guida
4:00 - 4:15	Speaker 16	FF0048	TE-Tracker: systematic identification of transposition events through whole-genome resequencing	Mohammed-Amin Madoui
4:15 - 4:30	Speaker 17	FF0035	Insights from analyzing *Clostridium botulinum* sequences	Karen Hill
4:30 - 4:45	Speaker 18	FF0037	AntibodyMining ToolBox: An Open Source Tool for the Rapid Analysis of Antibody Repertoires	Csaba Kiss
4:45 - 5:00	Speaker 19	FF0077	Genome evolution and GC patterns driven by recombination	Anitha Sundararajan
5:00 - 5:15	Speaker 20	FF0067	Comprehensive identification of structural variants in a robustly characterized personal human genome	William J Salerno
5:15 - 5:30	Speaker 21	FF0079	Methyl Sequencing, Dissecting the Subtleties of the Differentiated and Un-differentiated Genome	Masoud Toloue
6:00 - 8:00pm	Happy Hour(s)	x	Happy Hour at Cowgirl Cafe - Sponsored by illumina - Map Will be Provided	Sponsored by illumina
8:00 - bedtime	on your own	x	Dinner and Night on Your Own - Enjoy!!!	x

NOTES

Speaker Presentations (May 29th)

Abstracts are in order of presentation according to Agenda

Keynote

FF0111

Dissecting the Missing Diagnostic Yield in Exome sequencing

Deanna Church

Personalis, Inc.

Whole exome sequencing is playing a bigger role in research and clinical testing. During these early days it is critical that we examine the entire testing process to find ways to improve exome tests from current published estimates of 25% diagnostic yield. The complex process of going from DNA sample to clinical report involves multiple, technologically complex steps including sample prep, sequencing, alignment and variant calling, annotation, and interpretation, each of which can be improved to increase the chances of finding a causative variant. In this talk I will specifically address issues affecting variant identification as well as interpretation.

NOTES

FF0057

A *de novo* Whole Genome Shotgun Assembler for Long Reads

Gene Myers, Director & Tschira Chair

Max Planck Institute for Molecular Cell Biology and Genetics

10Kbp long reads are wonderful but with a 15% error rate they are hard to work with. However, truly random error position and nearly Poisson single-molecule sampling imply that in principle reference quality reconstructions of gigabase genomes are possible with as little as 30X coverage. Such a capability would resurrect the production of true reference genomes and enhance comparative genomics, diversity studies, and our understanding of structural variations within a population.

We are building an assembler we call the Dazzler (the Dresden AZZembLER) [1] that can assemble 1-10Gb genomes *directly* from a shotgun, long read data set currently producible only with the PacBio RS II sequencer. It is based on the string graph paradigm [2] with its two most important attributes being:

1. It reasonably scales to gigabase genomes being roughly 35X faster than current assemblers [3] for this kind of data.
2. Using the relationship of a given read to all other reads, we carefully identify artifacts in a read, estimate the quality of small stretches of a read, and use these implied qualities to correct a read's sequence more accurately then in previous approaches [3].

We will report on our progress and results to date for several target genomes and data sets.

[1] www.dazzlerblog.wordpress.com.

[2] E. Myers, "The Fragment Assembly String Graph", *European Conf. on Computational Biology* (Madrid, Spain, 2005), 79-85.

[3] C.S. Chin, D.H. Alexander, P. Marks, A.A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E.E. Eichler, S.W. Turner, and J. Korlach, "Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data", *Nature Methods* 10 (2013), 563-569.

FF0044

Speeding up NGS software development

E. Drezen[1], G. Rizk[1], R. Chikhi[2], C. Deltel[1], C. Lemaitre[1], P. Peterlongo[1], D. Lavenier[1]

1 INRIA/IRISA/GenScale, Campus de Beaulieu, 35042 Rennes cedex, France
2Department of Computer Science and Engineering, Pennsylvania State University, USA

The analysis of NGS data remains a time and space-consuming task. Many efforts have been made to provide efficient data structures for indexing the terabytes of data generated by the fast sequencing machines (Suffix Array, Burrows-Wheeler transform, Bloom Filter, etc.). Mapper tools, genome assemblers, SNP callers, etc., make an intensive use of these data structures to keep their memory footprint as lower as possible.

The overall efficiency of NGS software is brought by a smart combination of how data are represented inside the computer memory and how they are processed through the available processing units inside a processor. Developing such software is thus a real challenge, as it requires a large spectrum of competences from high-level data structure and algorithm concepts to tiny details of implementation.

We have developed a C++ library, called GATB (Genomic Assembly and Analysis Tool Box) to speed up the design of NGS algorithms. This library offers a panel of high-level optimized building blocks. The underlying data structure is the de Bruijn graph, and the general parallelism model is multithreading. The GATB library targets standard computing resources such as current multicore processor (laptop computer, small server) with a few GB of memory. Hence, from high-level C++ API, NGS programing designers can rapidly elaborate their own software based on state-of-the-art algorithms and data structures of the domain.

To demonstrate the efficiency of the GATB library, several NGS software have been designed such as contiger (Minia), read corrector (Bloocoo) or SNP discovery (DiscoSNP). The GATB library is written in C++ and is available at the following web site http://gatb.inria.fr under the GNU Affero GPL license.

FF0038

New Frontiers of Genome Assembly with SPAdes 3.1

Anton Korobeynikov[1,2], Dmitry Antipov[1], Anton Bankevich[1], Alexey Gurevich[1], Sergey Nurk[1], Andrey D. Prjibelski[1], Yana Safonova[1], Irina Vasilinetc[1], Alla Lapidus[1,3] and Pavel Pevzner[1,4]

1 Algorithmic Biology Laboratory, St. Petersburg Academic University, St. Petersburg, Russia 2 Faculty of Mathematics and Mechanics, St. Petersburg State University, St. Petersburg, Russia 3 Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia 4 Department of Computer Science and Engineering, University of California, San Diego, USA

Despite all the efforts high quality genome assembly is a complex task that so far remains unsolved. It is well known that majority of problems caused by repeats present in all genomes of any nature. The usage of multiple methods of genomic DNA isolation, different sequencing technologies and different types of genomic libraries for research projects introduces additional levels of complication to the genome assembly. The assembler tool SPAdes was originally developed at the St. Petersburg Academic University for the purpose of overcoming the complications associated with single-cell microbial data (uneven coverage and increased level of chimerical reads). The tool was able to successfully resolve these issues for Illumina reads and was recognized by the scientific community as one of the best assemblers working with both isolates and single-cell data.

Current SPAdes 3.0 version of the assembler provides an ability to work with different combinations of sequencing platforms including Illumina, Ion Torrent, PacBio, Sanger using both paired-end and mate-pair libraries of different insert sizes. Here we present SPAdes 3.1 - further development of SPAdes, which resolves many scaling issues with respect to running time and RAM consumption as well as improved support for IonTorrent data, updated algorithms for scaffolding and repeat resolution, and an approach for mate-pair only assembly from the reads produced by novel Illumina NexteraMP protocol.

Keywords: de novo assembly, hybrid assembly, repeat resolution, scaffolding, mate pairs

FF0065

Upgrading large genomes using Pacific Biosciences long reads and PBJelly software

Jeffrey Rogers, Adam English, Yi Han, Stephen Richards, Muthuswamy Raveendran, Daniel Hughes, Vanessa Vee, Mark Wang, David Rio Deiros, Yue Liu, Viktoriya Korchina, Donna Muzny, Kim Worley and Richard Gibbs

Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX

Despite advances in genome re-sequencing, the *de novo* assembly of large polymorphic genomes (>2 Gb) remains a challenging problem. Next generation short read (100-250bp) sequencing is used with a number of different whole genome assembly tools to generate valuable draft quality assemblies. But contig N50 values for those assemblies rarely reach beyond 75kb. More recent strategies have used datasets consisting only of reads from the Pacific Biosciences RS II platform, with some success, but *de novo* assembly of mammalian (~3 Gb) diploid genomes has not yet been achieved with that strategy. Using PBJelly 2.0, we mapped 5-19x whole genome PacBio read coverage to draft genome assemblies produced using Illumina Hi-Seq reads and the ALLPATHS-LG assembler, followed by additional scaffolding and gap filling using Illumina reads and Atlas-Link and Atlas-GapFill. The latest PBJelly 2.0 software efficiently leverages the current PacBio SMRT analysis pipeline and adds scaffolding to fill inter-scaffold gaps as well as the intra-scaffold gaps addressed by PBJelly 1. Other improvements increased the accuracy of the gap-filling sequence. Using PBJelly2.0, we obtained the following contig N50 values for mammalian genomes: sheep 501 kb; rat 131 kb; baboon 376 kb; sooty mangabey 229 kb; mouse lemur 290 kb. This approach consistently produces *de novo* assemblies with contiguity sufficient to ensure that most genes fall into a single contig. Methods that only use PacBio reads without Illumina data require much higher coverage and thus are more expensive (5- to 10-fold more). The factors that influence ultimate contig N50 results obtained through our combined approach are under active investigation, but likely include the length of the PacBio reads, the size of the ALLPATHS + Atlas-Link scaffolds, level of heterozygosity of the genome and the number and distribution of repetitive elements.

Keywords: de novo assembly, PBJelly2.0, mammalian genomes, Pacific Biosciences

FF0014

Assembly and Phasing of Polymorphic Heterozygous Diploid Genomes

Jason Chin (jchin@pacificbiosciences.com)

Pacific Biosciences

Genome assembly for polyploid genomes remains a challenging task. The longer sequences generated from PacBio® RS II provide a new opportunity to generate better assemblies with longer contiguity for diploid genomes. The contiguity of a diploid assembly is a function of the degree of variation between homologous chromosomes. It is essential to create diploid genome sequence data sets and use them to guide the algorithm development work in order to provide better tools for diploid genome assembly for biological study. We will present diploid assembly results using various synthetic F1 Arabidopsis hybrids that have varying degrees of similarity between the homologous chromosomes. Inbred 'parental strains' were first sequenced and assembled individually to provide a phased reference for subsequent in-silico diploid assemblies. The sequence of these inbred strains were generated using long insert (>15 kb) SMRTbell libraries of inbred Arabidopsis lines and sequenced on the Pacific Biosciences RS II instrument with either P4-C2 or P5-C3 chemistries. Typically these inbred lines yielded raw sequence read length of ~ 9 kb (average) with maximum read lengths of ~35 kb. The raw reads (~60X coverage) of the inbred lines were assembled using our hierarchical genome assembly process (HGAP) yielding contig N50s > 5 Mb. By mixing two or more data sets (two or more strains) it is possible to create in-silico diploid data sets in which we know the exact sequence of the parental chromosomes. Additionally, mis-assemblies can easily be identified and addressed, as the origin of the reads is known. A new string graph based assembler was employed to assemble the reads. The impact of the degrees of variation on the assembly contiguity and phasing information are analyzed with different mixtures of data set to provide better understanding for the best strategy for assembling diploid genomes.

Keywords: heterozygous poly-ploid genome assembly

FF0063

De novo Assembly of *Medicago truncatula* Genome Lines Using Illumina and Pacific Biosciences Sequencing Technologies

Thiruvarangan Ramaraj[a], Brian Walenz[b], Nevin Young[c], Kevin Silverstein[c], Joann Mudge[a], Jason Miller[b]

[a]National Center for Genome Resourcess (NCGR), Santa Fe, NM, USA [b]J Craig Venter Institute (JCVI), Rockville, MD, USA [c]University of Minnesota, Minneapolis, MN, USA

Pacific Biosciences (PacBio), Moleculo, and Oxford Nanopore, among other third generation single molecule sequencing technologies, are revolutionizing research in the area of genomics and transcriptomics with their capability to sequence long individual molecules of DNA and RNA. One major advantage is the ability to generate longer sequences: thousands or tens of thousands of base pairs instead of mere hundreds. With the rapid advancements in second and third generation sequencing technologies, hybrid genome assembly approaches using the strengths of both of these technologies is gaining impetus. While the hybrid approaches have been shown to work well largely in the microbial world, recently there has been greater push to extend them beyond microbes and into larger and complex eukaryotes like plants, animals, and humans.

We will present preliminary results toward generating *de novo* genome assemblies for *Medicago truncatula* plant accessions using a multi-platform sequencing strategy (Illumina HiSeq & PacBio RSII). One of the key aspects of this study is to explore structural variation in the context of *Medicago truncatula* diversifications, requiring accurate long-range contiguity in the *de novo* reconstructions. We have developed a hybrid assembly pipeline to integrate HiSeq and PacBio RSII data. The pipeline exploits the base call accuracy and read pair linkage of Illumina data and the long read contiguity of PacBio data. It uses Celera Assembler to integrate corrected PacBio reads from ECTools with pseudo-mate-pairs sampled from ALLPATHS-LG scaffolds. This approach produced larger contigs and scaffolds than any of the constituent tools generated separately.

Keywords: Third Generation Sequencing, *Medicago truncatula,* Multi-platform Sequencing Approach, Hybrid Genome Assembly Workflow.

FF0062

Illumina sequencing with no artifacts

Zbyszek Otwinowski & Dominika Borek

UT Southwestern Medical Center in Dallas

We devised a novel, integrated method of data analysis that will provide accurate disentanglement of fluorescence signals, so that the nucleotide bases can be called with high certainty and assured statistical independence. We elucidated all the factors that contribute to sequencing errors and identified robust and efficient approaches to correct for them. To detect the underlying sources of problems, we start with generating sequencing data by PCR-based clonal amplification, which generates multiple copies of the same sequence. These copies are expected to have identical fluorescence signals, with exceptions resulting from sporadic PCR errors and instrumental effects (*e.g.* air bubbles). We have been able to explain the mechanisms of all the published artifacts of Illumina sequencing and we have also detected additional, unpublished sources of error. We implemented algorithms that correct for these error sources in the final results. Illumina produces images with a very high signal-to-noise ratio that can reach the level of 100:1 for the signal in one pixel. Therefore, problems in data analysis result from the need to keep a sufficiently high accuracy of approximations through all the steps rather than from the level of noise in the measurements. Our reference point – the standard Illumina pipeline – includes non-optimal approximations, while alternative base-callers address only a small fraction of the weaknesses in Illumina's pipeline. We designed our base-calling to progressively narrow the uncertainty of results through iterative procedures.

We tested these methods on multiple sequencing data sets and have been able to efficiently classify problems into distinct categories such as instrumental problems and chemistry that depends on either local or distal sequence contexts. These novel methods will positively affect all aspects of genomic sequencing.

FF0034

Near perfect de novo assemblies of eukaryotic genomes using PacBio long read sequencing

Gurtowski, J.[1], Deshpande, P.[1], Eskipehlivan, S. M.[1], Kramer, M.[1], Lee, H.[1] Goodwin, S.[1], Antoniou, E.[1], Heiner, C.[2], Khitrov, G.[2], Schatz M.C.[1] and McCombie, W. R.[1]

1) Cold Spring Harbor Laboratory, and 2) Pacific Biosciences,

Recent advances in long read sequencing technology have allowed the scope of assembly projects to increase dramatically. It is now possible to produce near perfect eukaryotic draft assemblies using Pacific Bioscience's SMRT sequencing technology. We recently sequenced multiple strains of yeast including both S. cerevisiae and S. pombe. When assembled with a combination of HGAP and the Celera assembler all of the chromosomes were present in a single contig or a very small number of contigs. In addition, we also compared assemblies from several species with larger genomes including Arabidopsis and a few strains of rice in which we achieved contig N50 sizes ranging from hundreds of thousands to millions of basepairs. The larger genomes have lead us to explore different hybrid error correction approaches and we present our algorithms and analyses. We have found that hybrid correction can produce assemblies with contig sizes similar to, but at a fraction of the cost of, self-correction.

Keywords: genome assembly, near perfect, hybrid correction

NOTES

NOTES

Lunch

12:15 – 1:55pm

Sponsored by

Notes

FF0069

Anchored Assembly: An Algorithm for Large Structural Variant Detection Using NGS Data

Jeremy Bruestle, Becky Drees, and Niranjan Shekar

Spiral Genetics, Seattle, WA

Characterization of large indels, inversions, and multi-nucleotide variants is important for disease, agrigenomics, and microbial genomics studies. These are often undetected by standard pipelines. Spiral Genetics has developed Anchored Assembly, a novel method using direct, *de novo* read overlap assembly to accurately detect variants from next-generation sequence reads. We detect, on average, over 90% of indels and structural variants up to 30 kbp in non-repetitive regions. The ability to detect deletions and structural variants is undiminished by variant size, and the ability to accurately detect and assemble insertions continues well into the 30 kbp range.

Anchored Assembly was evaluated against Pindel and BWA + GATK using simulated read data. Datasets were generated by populating chromosome 22 of the human genome reference sequence with a set of SNPs, insertions, deletions, inversions, and tandem repeats. Overall, Anchored Assembly detected over 90% of indels and structural variants up to 50 kbp and SNPs with false discovery rates well below 1%. In comparison, Pindel and BWA + GATK had overall false discovery rates of 10% and 9%, respectively.

Anchored Assembly's range of detection and low false discovery rates may be useful for characterizing tumor vs. normal samples and analyzing and comparing bacterial strains.

Keywords: structural variation, genome assembly algorithms

FF0085

De Novo Assembly and Structural Analysis Using Extremely Long Single-Molecule Imaging

H VanSteenhouse, A Hastie, K Haden, Z Dzakula, M Austin, F Trintchouk, M Saghbini, H Cao

De novo genome assemblies built based on only short read data are generally incomplete and highly fragmented due to the intractable complexity found in most genomes. This complexity, consisting mainly of large duplications and repetitive regions, hinders sequence assembly and subsequent comparative analyses.

The Irys platform from BioNano Genomics overcomes the limitations of short fragment technologies to provide unprecedented insights into whole-genome biology. Irys is a single-molecule genome analysis system based on NanoChannel Array technology that linearizes extremely long DNA molecules for observation. This high-throughput platform automates massively parallel imaging of individual molecules of genomic DNA hundreds of kilobases in size to provide sufficient sequence uniqueness and long-range contiguity critical for unambiguous *de novo* assembly of complex genomes. High-resolution genome maps assembled *de novo* retain the original context and architecture of the genome, making them extremely useful for sequence assembly scaffolding and structural variation detection applications.

These genome maps provide dense genome-wide anchor points for ordering and orienting sequencing contigs or scaffolds to greatly increase completeness and accuracy of *de novo* assemblies. Structural variants and repeats are measured directly within long “reads” for comprehensive analysis of what has been dubbed “the inaccessible genome”. Irys genome maps have been constructed for the assembly and characterization of numerous genomes, including human, plant, fungi, and bacteria. Following an introduction to the platform and underlying single-molecule technology, several real-world examples will demonstrate its application to large complex genomes.

FF0020

Human sequence assembly scaffolding using Irys genome maps

H.Dai[1], A. Pang[1], A. Hastie[1], W. Stedman[1], Z. Dzakula[1], P- Y. Kwok[2], A. Ummat[3], A. Bashir[3], H. Cao[1]
1. BioNano Genomics, San Diego, CA
2. Mount Sinai School of Medicine, New York, NY
3. University of California, San Francisco, CA

Genome mapping technology from BioNano Genomics provides a platform for direct analysis of extremely long genomic DNA (up to multi-megabases) without amplification. *De Novo* assembly of these single molecule images can yield high fidelity contiguous information across long ranges, particularly in highly repeated regions. Resulting genome maps thus greatly complement assemblies using relatively short second- and third-generation sequencing reads.

We have collected 90-fold depth coverage of the human NA12878 sample from the CEU trio and constructed *de novo* consensus genome maps with N50 length of 4.6 Mb. Separately, a Pacific Biosciences sequence based assembly was produced for the same genome with N50 length of 930 kb. By combing data from these two technologies with a custom designed merging pipeline, we were able to generate an assembly having scaffold N50 length of greater than 10 Mb covering more than 2.7 Gb of the human genome. At the same time, we were able to identify potential misassembles by reviewing the inconsistencies between these two complementary technologies.

FF0071

Efficient de novo assembly of long NGS reads

Martin Simonsen, Poul Liboriussen, Kasper G. Larsen, Aske S. Christensen and Marta Matvienko.

Qiagen Aarhus, Denmark

As long NGS reads from platforms such as PacBio, Moleculo and Oxford Nanopore are becoming widely available, there is a growing need for efficient and user friendly tools that can assemble these reads. The error profile of the long NGS reads produced by current technologies make it difficult, and in some cases impossible, to use a de Bruijn graph based approach to assemble the reads in a computationally efficient way. Consequently, computationally expensive methods based on the more traditional overlap assembly approach, are now being used again to assemble large amounts of NGS reads. The assemblies produced by overlap based methods, such as the Celera assembler, contain long contigs of a high quality, but these methods are also time consuming and require large amounts of memory to run.

We present two software tools that provide de novo assembly of long reads in a computational efficient way. The first tool is a scaffolder that given a set of existing contigs use long NGS reads to join neighboring contigs and fill out gaps, which can improve the assembly quality significantly. The tool can take advantage of both high quality reads such as Moleculo reads and raw PacBio reads. The second tool is a de novo assembler that apply a hybrid approach for assembling long NGS reads. A de Bruijn graph is used to create an initial assembly of the reads which is then iteratively improved using an overlap based approach. Both tools can be used on a laptop computer where they are able to create results in a matter of minutes.

Keywords: de novo assembly, PacBio, Scaffolding, long reads.

FF0047

de novo mammalian assembly of one-library PCR-free 250-base Illumina reads

Iain MacCallum, Ted Sharpe, Neil Weisenfeld, Carsten Russ, Jessica Alfoldi, Jeremy Johnson, Jason Turner-Maier, Broad Genomics Platform, Chad Nusbaum, Kerstin Lindblad-Toh, David B. Jaffe

Broad Institute, 320 Charles Street, Cambridge MA 02142

Comparison of mammalian genomes aids understanding of the human genome, however currently available genomes are expensive to produce and limited in resolution. Existing approaches for generating high-quality mammalian assemblies require either the construction of multiple, labor-intensive libraries or the use of data types having a high per-base cost. To enable sequencing of hundreds of mammalian genomes, data costs need to be reduced and algorithms need to be developed to fully exploit these new data types.

We set out to test the potential of single library assemblies using a new low-cost data type, 250-base paired-end Illumina reads from a PCR-free library. One HiSeq 2500 flowcell yields ~60x coverage, at a cost of ~$3,500 in list price reagents. These data have the power to resolve perfect repeats of size up to ~500 bp, and although most genomes have long perfect repeats, they are usually rare within a single mammalian genome. Because the reads are created from a PCR-free library, they are relatively unbiased and hence have high intrinsic contiguity.

To generate de novo assemblies using this new data type, we devised a new algorithm that assembles a 60x mammalian genome in approximately 24 hour on a 32-core server. Our algorithm greatly reduces the complexity of read error correction by first generating a draft assembly from uncorrected reads, and then using this initial assembly to guide targeted error correction. Our algorithm then builds a full, accurate assembly using the corrected data.

We demonstrate our method with assemblies for aardvark and white rhino, two mammals that have been previously studied using more expensive techniques. We assess each assembly in comparison to the reference genome for the species and demonstrate the power of this technique for comparative genomics through comparison of these assemblies to the human genome.

FF0032

Creating a Single Haplotype Human Genome Assembly

Tina Graves-Lindsay, Robert Fulton, Karyn Meltz Steinberg, Wes Warren, Richard Wilson

The Genome Institute at Washington University School of Medicine

The human genome reference sequence has provided a foundation for studies of genome structure, human variation, evolutionary biology and human disease. At the time the reference genome was originally completed, it was clear, that there were some loci recalcitrant to closure with the technology and resources available. What was not clear was the degree to which structural variation and diversity affected our ability to produce a representative genome sequence at these loci. Many of these regions in the genome are associated with large, repetitive sequences. They exhibit complex allelic diversity such that de-convoluting these regions with DNA from a single donor can be complicated. In order to eliminate the complications of multiple alleles, we have utilized DNA from a hydatidiform mole, CHM1, which is essentially haploid. To achieve a single allelic representation of the entire CHM1 genome, we have generated ~200X whole genome shotgun sequence as Illumina paired end data, as well as over 469 BAC sequences from the CHM1 BAC library, CHORI-17. The whole genome data was assembled using a reference-guided assembly and the finished BAC sequences were incorporated into this assembly. Currently the assembly has the longest N50 contig length, >140,000bp, of any other human whole genome assembly submitted to GenBank but the goal for this genome is to produce an assembly that is the same quality as GRCh38, the Human Reference Assembly. To achieve this, we aim to incorporate next generation sequencing technologies, such as Pacific Biosciences, to increase the contig N50 and in turn, the completeness of the genome. A BioNano genome map has been created, which will aid in increasing the scaffold length and the quality assessment of the assembly. By using these new technologies in conjunction with the existing data, we expect to achieve our goal of a complete single haplotype human genome assembly.

FF0033

Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)

Stephanie M. Guida, Kelly B. Schilling, Connor T. Cameron, Peter B. Ngam, Jennifer L. Jacobi, Pooja E. Umale, Ken A. Seal, Mary K. Myers, Robin S. Kramer, Arvind K. Bharti, John A. Crow, and Callum J. Bell

National Center for Genome Resources (NCGR)

Marine microbial eukaryotes comprise a vast array of single-celled, nucleated microbes, including diatoms, dinoflagellates, amoebae, ciliates and water molds. These organisms fill numerous ecological roles ranging from photosynthetic primary producers (base of aquatic food webs) to heterotrophic consumers of pre-formed organic compounds. Despite their small size, they are responsible for generating up to half of the world's oxygen.

Even though these organisms are abundant and ecologically important, the gene content of oceanic microbial eukaryotes has not been studied extensively. Therefore one of the major goals of this G.B. Moore Foundation funded project is to generate a catalog of expressed genes from ~700 such microbes. 678 samples representing 397 unique strains have been approved for inclusion in the project. All approved samples have had TruSeq libraries sequenced (Illumina Hi-Seq 2000 2x50-nt paired-end sequences) and transcriptomes assembled using a *de bruijn* graph based approach. Functional annotation was carried out based on searches against the PFam, SUPERFAMILY and TIGRFAMs HMM libraries as well as the UniProtKB/Swiss-Prot protein database. Raw sequence reads, transcriptome assemblies, annotations and metadata for 611 data sets are now publically available through CAMERA (http://camera.calit2.net/mmetsp/). The primary focus of this project is to increase the research community's scientific knowledgebase and enable metagenomic analyses of complex marine communities. This will lead to a better understanding of the ecology and evolution of these highly diverse organisms.

FF0048

TE-Tracker: systematic identification of transposition events through whole-genome resequencing

Arthur Gilly1,*, Mathilde Etcheverry4,5,6,*, Mohammed-Amin Madoui1,*, Julie Guy1, Antoine Martin4,5,6, Tony Heitkam4, Stefan Engelen1, Karine Labadie1, Jeremie Le Pen4,5,6,† , Patrick Wincker1,2,3, Vincent Colot4,5,6,‡ and Jean-Marc Aury1,‡

1 Commissariat a l'Energie Atomique (CEA), Institut de Genomique (IG), Genoscope, 2 rue Gaston Crémieux, BP5706, 91057 Evry, France 2 Centre National de Recherche Scientifique (CNRS), UMR 8030, CP5706, Evry, France 3 Universite d'Evry, UMR 8030, CP5706, Evry, France 4 Institut de Biologie de l'Ecole Normale Supérieure, F-75230 Paris Cedex 05, France 5 Centre National de la Recherche Scientifique (CNRS), UMR 8197, F-75230 Paris Cedex 05, France 6 Institut national de la santé et de la recherche médicale (INSERM) U1024, F-75230 Paris Cedex 05, France

* These authors contributed equally † Current address: Gurdon Institute and Department of Biochemistry, University of Cambridge, The Henry Wellcome Building of Cancer and Developmental Biology, Tennis Court Rd, Cambridge CB2 1QN, UK. ‡ Corresponding authors

Dedicated software for the detection of transposable element (TE) mobilization using next-generationsequencing is rare and often restricted to certain use-cases. Here, we present TE-Tracker, a general computational method for accurately detecting both the identity and destination of newly mobilized TEs from re-sequenced genomes. We perform a complete evaluation of existing software and show that TETracker, while working independently of any prior annotation, bridges the gap between generic structural variant (SV) detection tools and specialized TE-detection software, both in terms of specificity and sensitivity. We use TE-Tracker to provide a comprehensive view of transposition events induced by loss of DNA methylation in Arabidopsis. TE-Tracker is freely available at http://www.genoscope.cns.fr/TETracker.

FF0035

Insights from analyzing *Clostridium botulinum* sequences

Karen Hill

Bioscience Division, Los Alamos National Laboratory

Clostridium botulinum is a group of anaerobic spore-forming bacteria that share the common characteristic of expressing one or more of the seven botulinum neurotoxins designated BoNT/A-G. The species is genetically diverse by 16S *rrn* comparisons and could be considered as six different species.
Since the first completed genome of a BoNT/A-producing strain was released in 2007, the sequences of 25 other *C. botulinum* strains have been released. Because *C. botulinum* consists of seven BoNTs expressed in six species (designated as Groups) the genomes have greatly increased the understanding of the diversity of the toxin and the host bacteria.

Analysis of the genomes has shown that the *bont* gene can be within the chromosome, plasmid or phage and that its location does not appear to be random. The *bont* gene is associated with other genes within a toxin complex (~16 kb) and the completed genomes have provided an opportunity to examine the components of the toxin complex and its flanking regions. The availability of the genomic sequences has revealed the mobility of the *bont* gene and significant recombination events that have resulted in genetic variation.

FF0037

AntibodyMining ToolBox: An Open Source Tool for the Rapid Analysis of Antibody Repertoires

Csaba Kiss

Bioenergy and Biome Sciences, Los Alamos National Laboratory

Deep sequencing is an alternative method to analyze antibody repertoires and selections. The analysis provides an extremely detailed view of the selected antibody population and allows the identification of specific antibodies using only sequencing data. I will describe the development of AntibodyMining Toolbox, an open source software package for the straightforward analysis of antibody libraries sequenced by all three popular next generation sequencing platforms (454, Ion Torrent, MiSeq).

FF0077

Genome evolution and GC patterns driven by recombination

Anitha Sundararajan, Joann Mudge, Thiruvarangan Ramaraj, Madeline Kwicklis, Nathan Garcia, Oliver Oviedo, Kayla Engstrom, Michael Gonzales
UMN: Changbin Chen, Stefanie Dukowic-Schulze USDA, UMN: Shahryar Kianian, Penny Kianian Cornell: Wojtek Pawlowski, Yan He, Jaroslaw Pillardy, Qi Sun, Minghui Wang

NGCR, Santa Fe, NM

Nucleotide bias at the wobble position or the third codon position is a well-documented phenomenon both in eukaryotes and prokaryotes. In *Zea mays* and other monocots, the GC content in the third codon position of genes (GC3) shows a bimodal distribution. The two defined GC3 peaks in monocots experience divergent evolutionary pressure and differ functionally, with different classes of genes in each peak. High GC3 genes within monocots were mainly categorized into functions involving electron transport or energy pathways, response to biotic and abiotic stressors and signal transduction. These observations have been further corroborated in our studies.

Recombination has been perceived to be a driving force for increases in GC patterns. Here we further explore the relationship between recombination and GC content in each of the three codon positions (GC1, GC2, and GC3, collectively termed GCx) by comparing GCx content to motifs underlying double strand break hotspots and meiocyte-specific gene expression.

Our results indicate a high overlap between high GCx genes of all three categories and genes that contain the double strand break hotspot motifs, which is expected due to the GC rich nature of the hotspot motifs. On the other hand, genes that are important in meiosis (early prophase I) are biased against both high GCx genes and genes with the motif, thereby suggesting a possible role in preventing double strand breaks from occurring in important meiotic genes. The presence of GC and GCx bimodality in maize cannot be fully explained by the codon wobble hypothesis, suggesting another mechanism is driving the biomodality. As the precursors to gene conversion, crossing over, understanding double strand breaks including their locations and surroundings is critical to understanding the process of recombination in entirety.

FF0067

Comprehensive identification of structural variants in a robustly characterized personal human genome

Adam C English[1], Claudia Gonzaga-Jauregui[2], Oliver A Hampton[1], Shruthi Ambreth[1], Deborah Ritter[1], Simon White[1], Caleb Davis[1], Pamela Mishra[1], Christine R Beck[2], Mahmoud Dahdouli[1], Narayanan Veeraraghavan[1], Alicia Hawes[1], David A Wheeler[1], Donna M Muzny[1], Jeffrey Rogers[1], Kim C Worley[1], Aniko Sabo[1], William J Salerno[1], James R Lupski[2], Richard A Gibbs[1]

[1]Baylor College of Medicine, Human Genome Sequencing Center, Houston, TX, [2]Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX

The human genome is subject to large (>75 bp) deletions, insertions, inversions, translocations, and complex combinations of these events. Such structural variants (SVs) directly impact gene function and regulation and, in turn, human health and disease. To characterize the SV content of individual genomes, we have devised a novel approach for integrating whole-genome array comparative genomic hybridization (aCGH) and short-read next-generation sequencing (NGS) data with long-read (Pacific BioSciences RSII), long-insert (Illumina Nextera), and whole-genome architecture (BioNano Irys) data. These methods are illustrated via analysis of a preferred “reference genome”: a single phenotypically well-characterized individual with autosomal recessive Charcot-Marie-Tooth neuropathy. Analyses were performed with *Parliament*, a consensus SV-calling infrastructure that integrates existing and novel SV-detection methods, reduces their outputs, and annotates the results with genomic features and known variants. In this individual, *Parliament* identified 3,262 putative SV loci supported by at least two data types, 1,337 loci identified by multiple NGS methods, and 164 loci supported by multiple non-NGS data, a total of 4,763 loci covering approximately 1.5% of the genome. Subsequent hybrid local assembly successfully validated 3,155 (66.2%) of these events. By incorporating orthogonal data types, we characterized multiple complex events, including a single 40 Kbp chimeric event on chromosome 11 (p15.5) that NGS-only methods classify as independent inversion, deletion, and insertion calls. The *Parliament* infrastructure also accommodates data from the subject’s family members, SNP concordance, Sanger-validated loci, and copy-number variants from seven whole-exome technical replicates of the individual, which allowed us to compare 15 exonic SV loci identified from whole-genome and whole-exome data. These analyses suggest experimental guidelines to detect specific classes of SVs and a computational methodology for maximizing SV-detection accuracy. *Parliament* incorporates these guidelines into a whole-genome NGS BAM-to-SV pipeline implemented on the cloud-based service DNAnexus.

Keywords: Structural variation, long-read sequencing, hybrid assembly

FF0079

Methyl Sequencing, Dissecting the Subtleties of the Differentiated and Un-differentiated Genome

Masoud Toloue, (mtoloue@biooscientific.com)

Bioo Scientific

Methylation of cytosine (5-methylcytosine; 5mC) is a common epigenetic feature, pervasive in our genomes and fundamental for cellular differentiation processes and transcriptional control. Our knowledge about genome wide distribution of DNA methylation, how it changes during cellular differentiation and how it relates to chromatin modifications is limited. The role of DNA methylation in human disease has sparked interest in developing genome scale methods for DNA methylation profiling. Next generation DNA sequencing technologies have transformed the landscape of genomic research with their ability to produce gigabases of data in a single run. This has enabled researchers to perform genome wide and high depth sequencing studies that would normally not be possible. We will present our latest ChIP and Methyl-Seq data on genome wide comparisons between differentiated and un-differentiated tissues, explain the significance of these patterns and how to choose between methylated DNA immunoprecipitation, methyl-binding domains, and restriction enzyme based reduced methyl sequencing to get there. We will also describe the genetic basis for iPSC's somatic memory and reprogramming variability and how histone modification plays a leading role.

Keywords: Whole genome and reduced representation methyl sequencing

Notes

Happy Hour(s)

Cowgirl BBQ

505.982.2565   319 S. Guadalupe St   Santa Fe, NM

See map on next page!

6:00pm – 8:00pm, May 29th

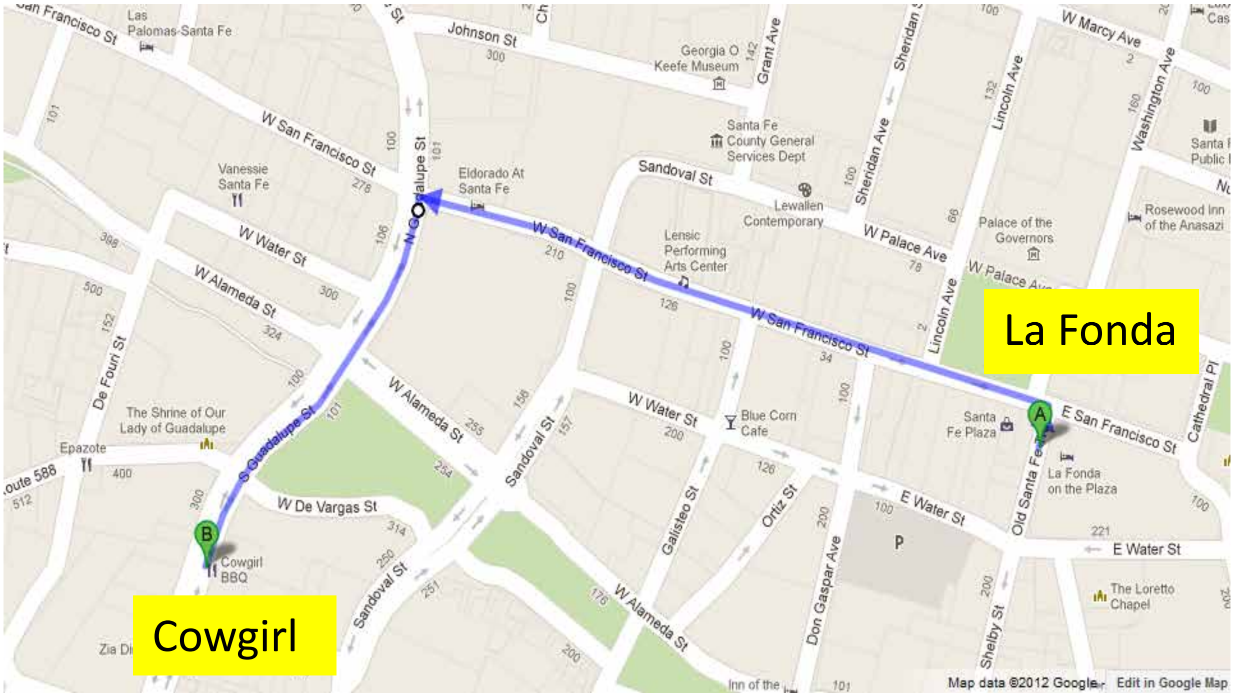
Drink tickets (margaritas, beer, sodas) will be provided

Sponsored by illumina!!!

illumina®

Map to Cowgirl BBQ

505.982.2565   319 S. Guadalupe St   Santa Fe, NM



Total Walking Distance

0.5 miles, 10 minutes

The Legend...

Many years ago, when the cattle roamed free and Cowpokes and Cowgirls rode the range, a sassy young Cowgirl figured out that she could have as much fun smokin' meats and baking fine confections as she could bustin' broncs and rounding up outlaws. So she pulled into the fine bustling city of Santa Fe and noticed that nobody in town was making Barbeque the way she learned out on the range. She built herself a Texas-style barbecue pit and soon enough the sweet and pungent scent of mesquite smoke was wafting down Guadalupe street and within no time at all folks from far and near were lining up for heaping portions of tender mesquite-smoked brisket, ribs and chicken. Never one to sit on her laurels, our intrepid Cowgirl figured out that all those folks chowing down on her now-famous BBQ need something to wash it all down with. Remembering a long-forgotten recipe from the fabled beaches of Mexico, she began making the now-legendary Frozen Margarita and the rest, as we say, is History. Before you could say "Tequila!" the musicians were out playing on the Cowgirl Patio and the party was in full swing.

05/30/2014 - Friday				
Time	**Type**	**Abstract #**	**Title**	**Speaker**
7:30 - 8:30am	**Breakfast**	**x**	**Harvey House Breakfast Buffet**	**Sponsored by NEB**
8:30 - 8:45	Intro	**x**	Welcome Intro from Los Alamos National Laboratory	**TBD**
x	Session Chair	**x**	Session Chairs	Chair - Patrick Chain Chair - Alla Lapidus
8:45 - 9:30	**Keynote**	**FF0112**	**The fly biome: Dispersal of Human Pathogens by Airborne Mechanical Vectors**	**Stephan Schuster** **Sponsored by Advanced Analytic**
9:30 - 10:00	Speaker 1	**FF0006**	Multigenerational exposure to the Chernobyl environment in bank voles alters the mitochondrial genome	**Robert Baker**
10:00 - 10:15	Speaker 2	**FF0051**	Functional profiling of genomic fragments with Sequedex	**Ben McMahon**
10:15 - 10:30	Speaker 3	**FF0003**	Rare OTUs Reveal Oral Microbiome Seeding Relationship between Tsimane Mother-Infant Dyads Who Practice Premastication	**Joe Alcock**
10:30 - 11:00am	**Break**	**x**	**Beverages and Snacks Provided**	**Sponsored by Kapa Biosystems**
11:00 - 11:15	Speaker 4	**FF0053**	Optimization of Metagenomic Methods for TEDDY Microbiome Study	**Ginger Metcalf**
11:15 - 11:30	Speaker 5	**FF0002**	Assessing the sensitivity of viral metagenomics	**Nadim Ajami**
11:30 - 11:45	Speaker 6	**FF0061**	Viral Metagenome Pipeline	**Christian Olsen**
11:45 - 12:00	Speaker 7	**FF0091**	Recruiting human microbiome shotgun data to site-specific reference genomes	**Gary Xie**
12:00 - 12:15	Speaker 8	**FF0028**	Analysis of Mixtures Using Next Generation Sequencing (NGS) of Mitochondrial DNA: Forensic Applications	**Henry Erlich**
12:15 - 12:30	Speaker 9	**FF0075**	Selective Depletion of Abundant RNAs to Enable Transcriptome Analysis of Low Input and Highly Degraded RNA from FFPE Breast Cancer Samples	**Bradley Langhorst**
12:30 - 1:45pm	**Lunch**	**x**	**Santa Fe Deli Lunch Buffet**	**Sponsored by MRI**
x	Session Chair	**x**	Session Chairs	Chair - Mike Fitzgerald Chair - Darren Grafham
1:45 - 2:00	Speaker 10	**FF0029**	A Targeted Sequencing Approach to Enable Enhanced Sensitivity in Variant Detection	**Bob Fulton**
2:00 - 2:15	Speaker 11	**FF0058**	*De Novo* Mapping with Solid-State Detectors	**John Oliver**
2:15 - 2:30	Speaker 12	**FF0004**	Further improvements to Illumina library preparation from challenging samples	**Maryke Appel**
2:30 - 2:45	Speaker 13	**FF0095**	Technology advancements in large insert PacBio library construction for targeted sequencing	**Min Wang**
2:45 - 3:00	Speaker 14	**FF0089**	Tools of the trade: resolving repetitive and complex regions in genomes using next-generation sequencing technologies and manual genome finishing	**Aye Wollam**
3:00 - 3:15	**Closing Discussions**	**x**	**Closing Discussions for General Meeting - discuss next year's meeting………..Now go out and enjoy Santa Fe!**	**Chair - Chris Detter**

NOTES

Speaker Presentations (May 30th)

Abstracts are in order of presentation according to Agenda

Keynote

FF0112

The fly biome: Dispersal of Human Pathogens by Airborne Mechanical Vectors

Ana Carolina Martins Junqueira[1], Rikky Wenang Purbojati[1], Aakrosh Ratan[2], Daniela I. Drautz[1], Lynn P. Tomsho[2], John J. McGraw[2], Caylie Hake[2], Gabriel Lopes Centoducatte[3], Daniel Fernando Paulo[3], Ana M. L. Azeredo-Espin[3], Staffan Kjelleberg[1], Donald A. Bryant[2], Bodo Linz[2], Stephan C. Schuster[1]

[1]Singapore Centre on Environmental Life Sciences Engineering Nanyang Technical University, Singapore [2]Center for Comparative Genomics and Bioinformatics, Penn State University, University Park, USA [3]Lab. Genética e Evolução Animal, Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, Sao Paulo, Brazil.

Carrion flies historically have been implicated in the transmission of diseases due to their attraction to feces and decaying organic matter. In addition to their medical and forensic importance, carrion flies play important environmental roles in nutrient recycling and pollination. In total, 128 microbiomes and 55 metatranscriptomes of individual flies were analyzed from urban, rural and rainforest environments on two continents. Using eight lines of evidence, we identified members of the microbiome to the genus or species level. Within twelve sampled environments, urban flies carry a surprising diversity of pathogens, with 42.8% of the microbiome being composed of dangerous human pathogenic microorganisms, suggesting that the carrion fly microbiome is largely determined by the environment rather by the insect host. The detected microbial diversity included *Acinetobacter baumannii*, *Clostridium botulinum, Escherichia coli*, *Helicobacter sp., Myroides odoratimimus*, *Proteus mirabilis*, *Salmonella enterica*, *Vibrio cholera, Yersinia enterocolitica* and *Y. pestis.* Studying total RNA, we detect a total of 737 OTUs with 92.7% of them being identified at the species level. The microbiome diversity of four body parts shows that pathogenic microorganisms are not restricted to the gastrointestinal tract and that legs and wings play a major role in pathogen transmission. Strikingly, organisms from the external surfaces were shown to be cultivatable and to coincide with the most abundant taxa identified in the meta'omic analyses. Furthermore, in laboratory reconstruction experiments we demonstrate that viable organisms are transferred between landing sites almost entirely by the feet of this mechanical vector. Thus, we infer that microbial contamination of humans, animals and plants may occur by fly-borne transfer of viable etiological agents capable of producing infections. Monitoring the microbiomes of mechanical vectors in synanthropic, agricultural and natural environments could significantly improve risk assessment and contribute to disease control.

NOTES

FF0006

Multigenerational exposure to the Chernobyl environment in bank voles alters the mitochondrial genome

R. J. Baker[1], Caleb Phillips[1], Jeffery Wickliffe[2], Faisal Anwarali Khan[3], Sergey Gaschak[4], Kateryna Makova[5] and Ben Dickens[5,6].

[1]Department of Biological Sciences and Museum, Texas Tech University, Lubbock, Texas, USA. [2]Department of Global Environmental Health Services, Tulane University, New Orleans, Louisiana, USA. [3]Department of Zoology, Faculty of Resource Science and Technology, Universiti Malaysia Sarawak, Kota Samarahan, Sarawak, 94300, Malaysia [4]International Radioecology Laboratory, Kiev Region, Slavutych, Ukraine. [5]Department of Biology, Penn State University, Old Main, University Park, Pennsylvania, USA. [6]School of Science and Technology, Nottingham Trent University, Nottingham, UK.

Using the bank vole (*Myodes glareolus*) as a model system, we tested the hypothesis that 40 generations of exposure to the Chernobyl environment will not significantly alter the mtgenome. We compared mtgenomes from 131 individuals including populations from sites with radioactive contamination comparable to that present in Northern Ukraine before the April 26 meltdown to populations living in the most radioactive sites (Red Forest and Glyboke Lake) created by the meltdown. Using a variety of population genetic measures we found multiple statistical differences in our comparisons of the populations from contaminated and uncontaminated localities. Therefore, we rejected the hypothesis of no significant genetic effect from 40 generations of exposure to Chernobyl radiation. Contaminated localities were found to be more diverse than uncontaminated localities, exhibiting greater numbers of haplotypes, polymorphic loci, and increased genetic diversity, and these differences were not attributable to sample size artifacts. The observed pattern of diversity is contrary to predictions of a source-sink demographic scenario, and is consistent low-dose radiation producing an accelerated mutation rate. This study is the first to demonstrate this phenomenon. However, bank voles from contaminated localities are not distinguishable from those collected in uncontaminated localities, compatible with the hypothesis that the cost of living in the most contaminated site does not have profound health effects.

Keywords: chronic exposure, multigenerational exposure to radiation, Chernobyl, Ukraine, mtGenome

FF0051

Functional profiling of genomic fragments with Sequedex

Ben McMahon, Nick Hengartner, Judith Cohn, Mira Dimitrijevic, and Joel Berendzen

We have previously shown [1,2] that amino acid 10-mers can be used to rapidly (6 Gbp/hr) classify short genomic fragments according to both phylogeny and function across the entire tree of life using a data module that fits into RAM of a typical desktop computer (4, 8, 16, or 32 GB of RAM). In contrast to phylogenetic placement algorithms, which are relatively well-developed, functional assignment of reads is complicated by both the diversity of proteins that are typically responsible for particular cellular functions and the tendency of evolution to re-tool proteins for a variety of distinct functions with relatively minor changes to the amino acid sequence. As a result, several large-scale efforts at protein function classification exist, such as COG, PFAM, Kegg, GO, or SEED. Sequedex can subsume any classification scheme which can be communicated by example gene sequences for each category, such as the 962 SEED subsystems ennumerated in Ref [2]. With the widespread availability of metagenomes and transcriptomes from applications as diverse as environmental microbiology, cell differentiation, algal bioengineering, and human microbiomes, as well as synthetic data generated from completed genomes of model organisms, it is possible to rapidly assess the utility of functional classification schemes across the variety of applications. We present such an assessment for the SEED classification scheme.

[1] Joel Berendzen, William J Bruno, Judith D Cohn, Nicolas W Hengartner, Cheryl R Kuske, Benjamin H McMahon, Murray A Wolinsky and Gary Xie, "Rapid phylogenetic and functional classification of short genomic fragments with signature peptides", BMC Res. Rep. 5:460, (2012) www.biomedcentral.com/1756-0500/5/460/

[2] http://sequedex.readthedocs.org/en/latest/

FF0003

Rare OTUs Reveal Oral Microbiome Seeding Relationship between Tsimane Mother-Infant Dyads Who Practice Premastication

Cliff S. Han[1], Melanie Martin[2], Armand Dichosa[1], Ashlynn Daughton[1], Seth Frietze[3], Michael Gurven[2], Joe Alcock[4]

1 Los Alamos National Laboratory, Bioscience Division, 2 University of California Santa Barbara,Dept. Anthropology, 3 University of Northern Colorado, School of Biological Sciences, 4 The University of New Mexico, Dept. of Emergency Medicine

Premastication is an ancient practice of feeding an infant with food chewed by its mother or other family members. Premasticated food may increase infant dietary availability, while caregivers' saliva transferred in the process may confer additional immunological and nutritional benefits. It has been recently postulated that early childhood exposure to oral microbiota through premastictaion and other forms of contact may help prime the developing infant immune system and promote oral tolerance to food antigens. We studied the oral microbiota of Amerindian mother-infant pairs for evidence of oral microbial-transferring influenced by premastication. Infant and maternal salivary microbiomes were distinct..Similarity within a mother-infant dyad was lower than that among infants or among mothers at the community level. Analysis of rare operational taxonomy units (OTUs) indicated a possible seeding relationship within the mother-infant dyad, which may be due to premastication. Most infants were not colonized by select oral bacterial pathogens, especially red-complex bacteria, which were identified in mothers.

FF0053

Optimization of Metagenomic Methods for TEDDY Microbiome Study

Ginger A. Metcalf[1], Nadim J. Ajami[2,3], Harsha Doddapaneni[1], Michelle Bellair[1], Tulin Ayvaz[2,3], Lauren D. Railey[2,3], Tonya Bauch[2,3], Auriole Tamegnon[2,3], Matthew C. Wong[2,3], Daniel P. Smith[2,3], Diane S. Hutchinson[2,3], Matthew C. Ross[2,3], Yi Han[1], Eric Boerwinkle[4], Donna M. Muzny[1], Joseph F. Petrosino[1,2,3], Richard A. Gibbs[1], and the TEDDY Study Group

[1]Human Genome Sequencing Center, Baylor College of Medicine, [2]Alkek Center for Metagenomics and Microbiome Research, Baylor College of Medicine; [3]Department of Molecular Virology and Microbiology, Baylor College of Medicine, [4]Human Genetics Center, University of Texas Health Science Center at Houston

The Environmental Determinants of Diabetes in the Young (TEDDY) microbiome study represents the largest clinical microbiome study to date, surpassing the Human Microbiome Project in sample size and scope. This effort involves the DNA extraction, sequencing, and analysis of over 13,000 stool and 6,000 plasma samples. Optimizations to standard laboratory protocols were required in order to process this large number of samples with a greater than 95% success rate, while adhering to the study budget and timeline. Examples include the implementation of automated nucleic acid extraction protocols, modification of standard automated PE library preparation allowing for a reduction in DNA input requirements, and the implementation of a viral protocol that utilizes an internal barcoding scheme to allow library construction to be performed on a pool of samples reducing time and cost.

To date, WGS data has been generated on over 10,000 stool samples at an average sequencing yield of 1.2 Gb per sample, while 16S data (V4) has been generated on >13,000 stool samples. Early, limited analyses of the 16S rDNA and WGS data show age and geographic location have a significant correlation with observed alpha diversity within samples. Both 16S and WGS data show a shift in the community structure of the gut microbiome over time, characterizing deeply how the bacterial community changes from birth until stabilizing around age 3.

Data generation is currently underway for the other arms of the TEDDY microbiome study including WGS on plasma samples, viral metagenomic sequencing on stool and plasma samples, as well as 18S/ITS sequencing to profile eukaryotic microbial populations. Future analysis plans include use of extensive patient metadata to identify trends and potentially clinically relevant differences between case and control subjects. In addition, an integrated analysis is planned to incorporate other data sets available for the cohort including metabolomics, proteomics, and host gene expression data to provide more information about microbiome maturation and host genetic associations in individuals with T1D, and children in general.

FF0002

Assessing the sensitivity of viral metagenomics

Nadim J. Ajami[1,2], Matthew C. Ross[1,2], Matthew C. Wong[1,2], Elicia D. Grace[1,2], Ginger A. Metcalf[3], Donna M. Muzny[3], Richard A. Gibbs[3], TEDDY DCC[4], Richard E. Lloyd[2], and Joseph F. Petrosino[1,2,3]

[1]Alkek Center for Metagenomics and Microbiome Research (CMMR), [2]Department of Molecular Virology and Microbiology, and [3]Human Genome Sequencing Center (HGSC), at Baylor College of Medicine, Houston, TX; [4] TEDDY Data Coordinating Center, University of South Florida, Tampa, FL

Advances in next generation sequencing have enabled the field of virology to bypass cell-culture based techniques for the identification and characterization of viruses. The CMMR and the HGSC together are pursuing studies examining the role of the virome in Type I Diabetes (T1D). The Environmental Determinants of Diabetes in the Young (TEDDY) and the Network for Pancreatic Organ Donors with Diabetes (nPOD) cohorts enable us to detect the possible triggers for T1D onset through the analysis of over 20,000 samples.

We tested some of our latest strategies with an independent external quality assessment and proficiency testing organization, Quality Control for Molecular Diagnostics (QCMD), to determine the sensitivity of viral metagenomics (as compared to RT-PCR), in a blind trial for the detection of a panel of assorted enteroviruses (the suspected target for several of our studies). Twelve samples were delivered blindly to the CMMR containing one type of enterovirus (coxsackievirus, echovirus, or enterovirus) spiked with 10^3-10^6 viral copies per ml. After nucleic acid extraction, reverse transcription, and semi-random amplification, samples were sequenced on the Illumina HiSeq2500 platform. An average of 2.56GB of data (range: 0.06-4.7GB) were retrieved per sample. Sequencing reads were trimmed, demultiplexed, and low complexity filtered followed by subtraction of non-viral sequences. Nucleotide and translated nucleotide queries using Bowtie2 and USEARCH yielded positive alignments (70-99% sequence identity) to enteroviruses in nine out of the 11 positive samples. Of the two misidentified samples, one had suboptimal coverage (60MB) and the other had the lowest viral load ($2x10^3$ copies/ml).

Metagenomic approaches provide sensitive and agnostic means to profile microbiomes and describe novel disease etiologies. As we improve our methods, we rely on both metagenomics and targeted approaches, such as RT-PCR, to enhance our limits of detection. These data will shed further light on relevant microbiome/virome changes that occur during, and may be contributing to, the development of many diseases, including T1D

FF0061

Viral Metagenome Pipeline

Olsen, C.*[1], Qaadri, K.[1], Shearman, H.[2], Miller, H.[2], Ammundsen, B.[2]

[1] Biomatters, Inc. 60 Park Place, Suite 2100 Newark, NJ 07102 [2] Biomatters, Ltd. Level 2, 76 Anzac Ave. Auckland 1010 New Zealand

Metagenomic analysis is quickly gaining momentum as a well-suited technique to provide a detailed understanding into the composition and activity of bacterial and viral communities i.e. “microbiome”, “virome”. Understanding the role of the virome in health and disease requires a deeper understanding of their composition and dynamics under various environmental conditions within the human gut and other body sites. As the number of sequenced viromes increase, larger genomic fragments are resolved by assembling the large amount of sequence data generated for each metagenome. In this talk, we present a pipeline for virome analysis consisting of data preprocessing, assembly, annotation, and comparison. The metagenomic pipeline is able to utilize large data-sets comprising of viromes made of thousands of large genomic contigs. This pipeline can be used to analyze two types of datasets: (i) viromes composed of raw reads, mostly generated using 454 pyrosequencing technology and (ii) viromes assembled into contigs, a strategy made possible with datasets sequenced with either/both pyrosequencing or Illumina sequencing technologies. Users are able to explore and analyze viromes composed of raw reads or assembled fragments using Geneious R7, a user-friendly interface to explore any kind of virome and enables virologists to make the most of their metagenomics next generation sequence data.

FF0091

Recruiting human microbiome shotgun data to site-specific reference genomes

Xie G, Lo CC, Scholz M, Chain PS.

Los Alamos National Lab

The human body consists of innumerable multifaceted environments that predispose colonization by a number of distinct microbial communities, which play fundamental roles in human health and disease. In addition to community surveys and shotgun metagenomes that seek to explore the composition and diversity of these microbiomes, there are significant efforts to sequence reference microbial genomes from many body sites of healthy adults. To illustrate the utility of reference genomes when studying more complex metagenomes, we present a reference-based analysis of sequence reads generated from 55 shotgun metagenomes, selected from 5 major body sites, including 16 sub-sites. Interestingly, between 13% and 92% (62.3% average) of these shotgun reads were aligned to a then-complete list of 2780 reference genomes, including 1583 references for the human microbiome. However, no reference genome was universally found in all body sites. For any given metagenome, the body site-specific reference genomes, derived from the same body site as the sample, accounted for an average of 58.8% of the mapped reads. While different body sites did differ in abundant genera, proximal or symmetrical body sites were found to be most similar to one another. The extent of variation observed, both between individuals sampled within the same microenvironment, or at the same site within the same individual over time, calls into question comparative studies across individuals even if sampled at the same body site. This study illustrates the high utility of reference genomes and the need for further site-specific reference microbial genome sequencing, even within the already well-sampled human microbiome.

FF0028

Analysis of Mixtures Using Next Generation Sequencing (NGS) of Mitochondrial DNA: Forensic Applications

Hanna Kim[1], Daniela Cuenca, MS.[2], George Sensabaugh, D. Crim.[2,3], Henry Erlich, PhD.[1], Cassandra D. Calloway, PhD.[1,2]

[1]Children's Hospital & Research Center at Oakland; [2]University of California, Davis; [3]University of California, Berkeley

The analysis of forensics mixtures (samples that contain >one genotype) represents a significant technical and statistical challenge. The massively parallel and clonal aspects of next generation sequencing systems allows for "deep sequencing", the generation of thousands of clonal sequence reads, and "digital analysis", the resolution of the mixture by simply counting the number of reads corresponding to the component genotypes. Mitochondrial (mt)DNA is a valuable genetic marker for "deconvoluting" mixtures and for analysis of forensically relevant samples that are often limited and/or degraded. Typically, the component sequences (majority and minority type) differ by more than a single nucleotide, allowing the distinction of "signal" from "background noise". In addition, the high copy number of mtDNA genomes per cell increases the sensitivity. Finally, the haploid nature of the mtDNA genome means that each individual contributor is, in general, represented by a single mtDNA sequence, assuming that the issue of mtDNA heteroplasmy is addressed.

Using amplicon sequencing of the highly polymorphic hypervariable regions (HVI and II), artificially mixed samples were analyzed on the Roche 454 GS Junior instrument. The proportions, and the minority component sequences could be detected down to 1%. The assay is highly sensitive for sequencing limited DNA amounts as little as 0.5 pg genomic DNA input (~500 mtDNA copies). A system based on hybridization/probe capture was also developed to sequence the entire mtDNA genome and applied to the analysis of mixtures. Analyzing the entire mtDNA genome provides greater discrimination potential than the analysis of HVI/HVII. In addition, the use of hybridization probe capture is well suited to enrichment for the degraded mtDNA, found in many forensics specimens. These two systems for analyzing mtDNA sequence variation promise to be valuable approaches to the challenge posed by forensic mixed specimens.

FF0075

Selective Depletion of Abundant RNAs to Enable Transcriptome Analysis of Low Input and Highly Degraded RNA from FFPE Breast Cancer Samples

Bradley W. Langhorst

New England Biolabs, Inc

Deep sequencing of cDNA prepared from total RNA (RNA-seq) has become the method of choice for transcript profiling, and discovery. The standard whole-transcriptome approach faces a significant challenge as the vast majority of reads map to ribosomal RNA (rRNA). One solution is to enrich the sample RNA for polyadenylated transcripts using oligo (dT)-based affinity matrices; however, this also eliminates other biologically relevant RNA species, such as microRNAs and noncoding RNAs, and relies on having a high quality and quantity RNA sample.

Here, we present a method to eliminate abundant RNAs from total RNA with different degradation levels, from intact RNA to highly degraded formalin-fixed paraffin-embedded (FFPE) samples. This method is based on hybridization of probes to the targeted abundant RNA, followed by subsequent enzymatic degradation. We applied this method to remove cytoplasmic and mitochondrial rRNA from different eukaryotic total RNA samples (human, mouse and rat), as well as degraded (1 year old) and highly degraded (10 year old) FFPE breast tumor biopsy RNA samples. We evaluated the depletion efficiency and off target effect of this method using strand specific RNA high-throughput sequencing. Ribosomal RNA depletion resulted in a minimal percentage of total reads mapping to rRNA sequences, regardless of the species, input amount (1µg or 100 ng), or degradation level. Additionally, there was very good transcript expression (FPKM) correlation (>0.93) between rRNA depleted and non-depleted libraries.

This method offers a robust and simple solution for transcriptome analysis of a variety of samples, including low quality and low quantity clinical samples such as FFPE RNA. Moreover, it is amenable to high-throughput sample preparation and robotic automation. This method is sensitive, specific, and produces increased coverage of less abundant, non-targeted transcripts in RNA-Seq studies.

NOTES

Lunch

12:30 – 1:45 pm

Sponsored by

MRIGlobal

***National** Solutions* ***Worldwide** Impact*

Notes

FF0029

A Targeted Sequencing Approach to Enable Enhanced Sensitivity in Variant Detection

Robert S. Fulton, Timothy Ley, Richard K. Wilson and the Faculty and Staff of The Genome Institute at Washington University

As DNA sequencing efforts continue, it is clear that analyses are empowered by increased sequence depth. Depth improves detection sensitivity, variant allele frequency (VAF) precision, and increases statistical significance to findings. This is especially true in cancer samples, where tumor purity, clonality, and contamination, often compromise these detection capabilities. Deep digital read depth ameliorates these issues, but there is a significant cost associated with the generation of these data with >90Gb needed for WGS and >6-8Gb for exome, with coverage levels less than optimal for variation detection.

Here we present a hybrid capture approach using long (120bp) biotinylated oligonucleotides synthesized by Integrated DNA Technologies (IDT). These probes are provided in significant quantities, either pooled, or in individual tubes, offering flexibility in assay design and balancing efforts. As a stand-alone method, it provides deep digital data from highly multiplexed samples sets, for a fraction of the cost of whole genome or exome sequencing. Probes can also be coupled with an exome capture kit to enhance the coverage > 5X in the specific regions of interest, to more significantly empower variant discovery in that region without significantly impacting coverage in the remaining exome space.

A 264-gene probe set (1.4Mb) was designed from a list of recurrently mutated genes from acute myeloid leukemia (AML) cases as described in the TCGA AML marker publication (NEJM, May 2013). Additionally, other regions of interest from gene lists, viral targets, regulatory regions, and poorly covered regions of exome capture sets have been utilized, with great success. In this presentation, we show some of these efforts as well as outline potential future applications.

FF0058

De Novo Mapping with Solid-State Detectors

John S. Oliver, Brendan Galvin, Tony Forget, Sante Gnerre, Mike Kaiser, Peter Goldstein, Jennifer Davis, Matt Sooknah

Nabsys Inc

Genomes are hard to assemble: polymorphism, repeats, and sequencing bias can turn even small genomes into assembly nightmares. Long reads, however, can be used very effectively to inform the assembly process, both to disambiguate repeats, and to separate haplotypes of polyploid genomes.

Solid-state, electronic nanodetectors can generate long range mapping information from single molecules that are hundreds of kilobases in length. DNA is translocated through nanochannels and detected electronically. These molecules can be tagged at specific locations and those locations mapped at much higher resolution than is possible with optical methods. Reads are assembled with high efficiency and used to generate accurate reference maps for eukaryotic genomes. Long-range information is preserved so structural rearrangements and duplications are easily identified.

Examples of assemblies of phage, BAC, and bacterial genomes generated with a few hours of data collection will be shown. Currently, the instrument has 8 modules, each with a single nanodetector chip. The technology is highly scalable with the potential for much higher throughput by placing multiple detectors on each semiconductor-based chip, making analysis of large, complex genomes like human and plants feasible in the future.

FF0004

Further improvements to Illumina library preparation from challenging samples

Maryke Appel

KAPA Biosystems, Inc., Wilmington, MA

The expanding scope and application of next-generation sequencing in research and clinical environments fuels a demand for high-quality libraries from lower amounts of input DNA, and DNA of poor or variable quality. We previously reported on the benefits of evolved and optimally-formulated enzymes and a highly optimized “with-bead” protocol for the construction of DNA libraries for Illumina sequencing from challenging and low-input samples such as FFPE and ChIP DNA.

In ligation-mediated Illumina library construction, the percentage of input DNA that is successfully converted to sequenceable, adapter-ligated library decreases with decreasing input. With the “with-bead” KAPA Library Preparation Kit, conversion rates as high as 40% can be achieved for libraries constructed from 1 µg of high-quality DNA. This drops to 10 - 15% for 10 ng input, and to <1% for libraries prepared from high picogram quantities of DNA. In a recent collaborative study, conversion rates, rather than amplification efficiency, was also identified as the major bottleneck in library construction from FFPE samples. For both low-input and low-quality samples, low library yields limit library diversity, and result in high duplication rates and reduced coverage.

We have recently developed a streamlined DNA library construction kit with significant benefits for these challenging samples. The novel one-tube chemistry achieves conversion rates for FFPE DNA similar to those for high-quality samples. In addition, adapter-dimer formation is inhibited. As a result, library yields and diversity for low-input samples can be maximized by high adapter:insert molar ratios and longer ligation times.

Kapa’s evolved and optimally formulated enzymes, and lessons from DNA library construction have been applied in the development of our first kit for RNA library construction. Although the KAPA Stranded mRNA-Seq Library Preparation Kit is not specifically tailored for low-input or degraded RNA, it provides for improved coverage of GC-rich and low-abundance transcripts.

FF0095

Technology advancements in large insert PacBio library construction for targeted sequencing

Min Wang[1], Adam English[1], Christine Regina Beck[2], Yi Han[1], Fuli Yu[1], Eric Boerwinkle[1,3], James R. Lupski[2], Donna M. Muzny[1], Richard A. Gibbs[1]

[1]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, 77030, USA;[2]Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, 77030, USA;[3]Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, 77030, USA

Generation of long DNA sequence reads *without GC-bias is one unique property of PacBio* RS II system. Technology advancements have been made to utilize this property in human genome structural variation studies. *Currently in BCM-HGSC, small-size targets, e.g. haplotype phasing, are examined by long-range PCR-based amplicon sequencing. For medium-size, contiguous or multiple targets, we have* developed the PacBio Large Insert Targeted Sequencing (PacBio-LITS) method that integrates target enrichment technology into PacBio large-insert library preparation to facilitate structure variation delineation at specific chromosomal regions. To construct large-insert capture libraries with different insert sizes, we conducted a series of tests to optimize experimental conditions (e.g. DNA input, shearing, size selection, adaptor ligation, DNA polymerase activity) so as to make the protocol amenable to laboratory implementation. Enriched PacBio libraries with insert sizes of 6-8kb have been made from 1-2ug input DNA using custom probe sets and up to 70% target specificity has been achieved. The value of the new method is exemplified by delineation at breakpoint level of a low-copy-repeat (LCR)-associated complex genomic rearrangement (CGR) on chr17p11.2 in a patient diagnosed with the genomic disorder Potocki–Lupski syndrome (PTLS; MIM#610883). A total of 818 million aligned bases were generated from the 4.5-kb capture library using PacBio C2/XL chemistry, and ~70% of aligned reads mapped to the 7-Mb target region with the mean mapped sub-read length around 2kb. Identification of the 1bp microhomology-mediated inversion duplicate has led us to propose a DNA-replication based mechanism for the dup-normal-dup structural variant in the disease sample.

Keywords: targeted sequencing, single molecule sequencing, complex genomic rearrangement

FF0089

Tools of the trade: resolving repetitive and complex regions in genomes using next-generation sequencing technologies and manual genome finishing

Aye Wollam, Robert Fulton, Richard Wilson.

The Genome Institute at Washington University School of Medicine

Next-generation sequencing technology has revolutionized the era of genomics. Genomes are being sequenced at a rate never been done before and with high quality. Technology such as Illumina's sequencing by synthesis (SBS) allows sequencing to be done cheaply and massively. Single molecule real time (SMRT) sequencing technology developed by Pacific Biosciences has been shown to resolve highly repetitive genomic regions previously impervious to resolution by the first and second-generation DNA sequencing technologies. Genomes with regions of sequencing bias and complex structures may now be assembled with fewer gaps and at finish-grade quality without manual genome finishing. When the sequencing reads are longer than the largest repeats in the genomic region, the resulting assemblies are often contiguous with accurate repeat reconstruction. However, in regions of the genome with large segmental duplications, inverted repeats or large tandem arrays where the read lengths fall short of the repeat units, assemblies result in collapsed repeats and/or unsolved gaps. In these cases, manual review and editing of sequence data is necessary to correct misassemblies and to obtain an accurate representation of the sequence. Here, we describe the methods, strategies and tools that combine the different sequencing platforms with finishing protocols in order to resolve complex regions accurately and efficiently.

Keywords: genome finishing, next-generation sequencing, Illumina, Pacific Biosciences, misassemblies, accurate repeat reconstruction, strategies, tools

Notes

First Name	Last Name	Abstract #	Email	Institution
Nadim	Ajami	FF0002	nadim.ajami@bcm.edu	Baylor College of Medicine
Omayma	Al-Awar		oalawar@illumina.com	Illumina, Inc.
Joe	Alcock	FF0003	joalcock@salud.unm.edu	University of New Mexico
Laith	Al-Eitan		lneitan@just.edu.jo	Jordan Univ. of Science and Technology, Princess Haya Biotechnology Ctr
Johar	Ali		johar.ali@oicr.on.ca	Alviarmani
Robert	Allen		robert.w.allen@okstate.edu	Oklahoma State University
Areej	Al-Quran		Areejalquran@yahoo.com	Jordan Univ. of Science and Technology, Princess Haya Biotechnology Ctr
Brett	Ammundsen		brett@biomatters.com	Biomatters' Geneious
Cort	Anderson		cla@uidaho.edu	University of Idaho
Maryke	Appel	FF0004, FF0005	maryke.appel@kapabiosystems.com	Kapa Biosystems
Eugene	Arinaitwe	FF0107	earinaitwe@muienr.mak.ac.ug	Ugandan Ministry of Agriculture, Animal Industry and Fisheries
Jennifer	Ayres		jayrea@illumina.com	Illumina, Inc.
Melody	Baddoo		mbaddoo@tulane.edu	Tulane University
Marty	Badgett	FF0094	mbadgett@pacificbiosciences.com	Pacific Biosciences
Robert	Baker	FF0006	robert.baker@ttu.edu	Texas Tech University
Arpan	Bandyopadhyay		bandy016@umn.edu	University of University
Callum	Bell		cjb@ncgr.org	National Center for Genome Resources (NCGR)
Roby	Bhattacharyya	FF0007	rbhatt@broadinstitute.org	Broad Institute; MGH
Jonathan	Bingham	FF0008	binghamj@google.com	Google Inc.
Cecilie	Boysen		cboysen@clcbio.com	CLC bio - A Qiagen Company
Chris	Bradburne		Chris.Bradburne@jhuapl.edu	Johns Hopkins University, APL
Andrew	Bradbury		amb@lanl.gov	Los Alamos National Laboratory
Gavin	Braunstein		gavin.braunstein@dtra.mil	Defense Threat Reduction Agency
Raquel	Bromberg	FF0009	Raquel.Bromberg@utsouthwestern.edu	University of Texas Southwestern Medical Center at Dallas
David	Bruce		dbruce@lanl.gov	Los Alamos National Laboratory
Lijing	Bu		lijing@unm.edu	University of New Mexico
Sarah	Buddenborg		sbuddenb@unm.edu	University of New Mexico
Christian	Buhay	FF0010	cbuhay@bcm.edu	Baylor College of Medicine
Timothy	Byaruhanga	FF0011	tssekandi@gmail.com	National Influenza Centre, Uganda
Connor	Cameron		ctc@ncgr.org	National Center for Genome Resources (NCGR)
Han	Cao	FF0085	han@bionanogenomics.com	BioNano Genomics
Heather	Carleton-Romer	FF0012	hcarleton@cdc.gov	Centers for Disease Control and Prevention
Michael	Cassler		mcassler@mriglobal.org	MRI Global
Gvantsa	Chanturia	FF0013, FF0109	romail28@gmail.com	National Center for Disease Control and Public Health of Georgia
Olga	Chertkov		ochrtkv@lanl.gov	Los Alamos National Laboratory
Jason	Chin	FF0014, FF0015	jchin@pacificbiosciences.com	Pacific Biosciences
William	Chow	FF0016	wc2@sanger.ac.uk	Wellcome Trust Sanger Institute
Deanna	Church	FF0111	deanna.church@personalis.com	Personalis
Alicia	Clum	FF0017	aclum@lbl.gov	DOE Joint Genome Institute
Daniel	Colman		dcolman@unm.edu	University of New Mexico
Rebecca	Colman	FF0018	rcolman@tgen.org	Translational Genomics Research Institute
Sean	Conlan	FF0019	conlans@mail.nih.gov	National Human Genome Research Institute (NHGRI/NIH)
Luisa	Corredor		luisa.corredorarias@biofilm.montana.edu	Montana State University, Center for Biofilm Engineering
Helen	Cui	FF0108	hhcui@lanl.gov	Los Alamos National Laboratory
Heng	Dai	FF0020	hdai@bionanogenomics.com	BioNano Genomics
Hajnalka	Daligault		hajkis@lanl.gov	Los Alamos National Laboratory
Brent	Dalke		brent.dalke@thermofisher.com	ThermoFisher/IonTorrent
Karen	Davenport		kwdavenport@lanl.gov	Los Alamos National Laboratory
Matthew	Davenport		Matthew.Davenport@jhuapl.edu	Johns Hopkins University, APL
Ashley	DeAguero		amontoy8@unm.edu	University of New Mexico
Joseph	DeAguero		josephd@unm.edu	University of New Mexico
Nate	Dellinger		Nathan.Dellinger@noblis.org	Noblis
Xiang-yu	Deng		xdeng@uga.edu	University of Georgia
Vladimir	Dergachev	FF0021	vdergachev@bionanogenomics.com	BioNanoGenomics
Chris	Detter		cdetter@lanl.gov	Los Alamos National Laboratory
Nicholas	Devitt	FF0022	npd@ncgr.org	National Center for Genome Resources (NCGR)
Armand	Dichosa	FF0023	armand@lanl.gov	Los Alamos National Laboratory
Mira	Dimitrijevic		mira@lanl.gov	Los Alamos National Laboratory
Darrell	Dinwiddie	FF0024	dldinwiddie@salud.unm.edu	University of New Mexico
Norman	Doggett		doggett@lanl.gov	Los Alamos National Laboratory

Brinza	Dumitru		dumitru.brinza@thermofisher.com	ThermoFisher/IonTorrent
Colin	Dunn		Colin.Dunn@noblis.org	Noblis
Matthew	Dunn	FF0025	md3@sanger.ac.uk	Wellcome Trust Sanger Institute
Adam	English	FF0026	english@bcm.edu	Baylor College of Medicine
Tracy	Erkkila		terkkila@lanl.gov	Los Alamos National Laboratory
Henry	Erlich	FF0027, FF0028	herlich@chori.org	Children's Hospital & Research Center Oakland
Diego	Fajardo		dfajardo@ncgr.org	National Center for Genome Resources (NCGR)
Leigh	Fanning		leigh@versiera.net	Versiera
Jason	Farlow		jasonfarlow@hotmail.com	Farlow Scientific Consulting
Andrew	Farmer		adf@ncgr.org	National Center for Genome Resources (NCGR)
Michael	Farrell		mqf2@cdc.gov	Centers for Disease Control and Prevention
Kevin	Fengler		kevin.a.fengler@pioneer.com	Dupont Pioneer
Will	Fischer		wfischer@lanl.gov	Los Alamos National Laboratory
Haley	Fiske		hfiske@illumina.com	Illumina, Inc.
Michael	FitzGerald		fitz@broadinstitute.org	Broad Institute
Michael	Franklin		franklin@montana.edu	Montana State University
Charles	Fromer		charles.fromer@dtra.mil	Defense Threat Reduction Agency
Bob	Fulton	FF0029	bfulton@genome.wustl.edu	Washington University School of Medicine, The Genome Institute
Meredith	Gavin		gavin.meredith@thermofisher.com	ThermoFisher/IonTorrent
James	George		jgeorge@illumina.com	Illumina, Inc.
Lori	Gladney		hze1@cdc.gov	Centers for Disease Control and Prevention
Cheryl	Gleasner	FF0030	cdgle@lanl.gov	Los Alamos National Laboratory
Veena	Gnanakkan		fp.veena@gmail.com	Johns Hopkins University
Darren	Grafham	FF0031	darren.grafham@sch.nhs.uk	Sheffield Diag, UK
Tina	Graves-Lindsay	FF0032	tgraves@genome.wustl.edu	Washington University School of Medicine, The Genome Institute
Jane	Grimwood		jgrimwood@hudsonalpha.org	HudsonAlpha Institute
Stephanie	Guida	FF0033	sguida@ncgr.org	National Center for Genome Resources (NCGR)
James	Gurtowski	FF0034	gurtowsk@cshl.edu	Cold Spring Harbor Laboratory
Yazan	Haddad		yazanhaddad@hotmail.com	Jordan Univ. of Science and Technology, Princess Haya Biotechnology Ctr
Sandra	Halonen		shalonen@montana.edu	Montana State University
Cliff	Han		han_cliff@lanl.gov	Los Alamos National Laboratory
William	Hansen		william.hansen@thermofisher.com	ThermoFisher/IonTorrent
Kevin	Harrod		kharrod@lrri.org	Lovelace Respiratory Research Institute (LRRI)
John	Havens		jhavens@idtdna.com	Integrated DNA Technologies (IDT)
Matt	Hickenbotham	FF0104, FF0106	matthew.hickenbotham@thermofisher.com	ThermoFisher/IonTorrent
Karen	Hill	FF0035	khill@lanl.gov	Los Alamos National Laboratory
David	Hirschebrg		dlhirschberg@gmail.com	Columbia University
Kelly	Hoon	FF0113	khoon@illumina.com	Illumina, Inc.
Chris	Hopkins		vqd8@cdc.gov	Centers for Disease Control and Prevention
Andrew	Huang		wwm8@cdc.gov	Centers for Disease Control and Prevention
Mariela	Humphrey		mariela.humphrey@thermofisher.com	ThermoFisher/IonTorrent
Miriam	Hutchinson		yesterdaymail@gmail.com	University of New Mexico
Wei-Shou	Huu		wshu@umn.edu	University of Minnesota
Jennifer	Jacobi		jlj@ncgr.org	National Center for Genome Resources (NCGR)
Jonathan	Jacobs	FF0100	jjacobs@mriglobal.org	MRIGlobal
Ravi	Jain		ravi.jain@cbiocorp.com	cBio, Inc.
Xiaoben	Jiang	FF0036	sdpapet@unm.edu	University of New Mexico
Shannon	Johnson		shannonj@lanl.gov	Los Alamos National Laboratory
Tania	Jooste		Tania.Jooste@up.ac.za	University of Pretoria (Uganda)
Kyla	Jorgenson		kyla.jorgenson@okstate.edu	Oklahoma State University
Safak	Kalindamar		safakkalindamar@gmail.com	Mississipi State University
John	Kayiwa		jkayiwa@uvri.go.ug	Uganda Virus Research Institute
Jonathan	Kayondo		jkayondo@gmail.com	Uganda Virus Research Institute
Gladys	Kiggundu		gladyskiggundu@yahoo.com	Ugandan National Animal Disease and Epidemiology Center
Luke	Kingry		vtx8@cdc.gov	Columbus Tech and Services/Centers for Disease Control and Prevention
Csaba	Kiss	FF0037	csaba.kiss@lanl.gov	Los Alamos National Laboratory
Kristen	Knipe		wgg9@cdc.gov	Centers for Disease Control and Prevention
Gerwald	Koehler		gerwald.kohler@okstate.edu	Oklahoma State University
Bette	Korber	FF0092	btk@lanl.gov	Los Alamos National Laboratory
Anton	Korobeynikov	FF0038	anton@korobeynikov.info	St. Petersburg Academic University
Adam	Kotorashvili	FF0039	adam.kotorashvili@gmail.com	National Center for Disease Control and Public Health of Georgia

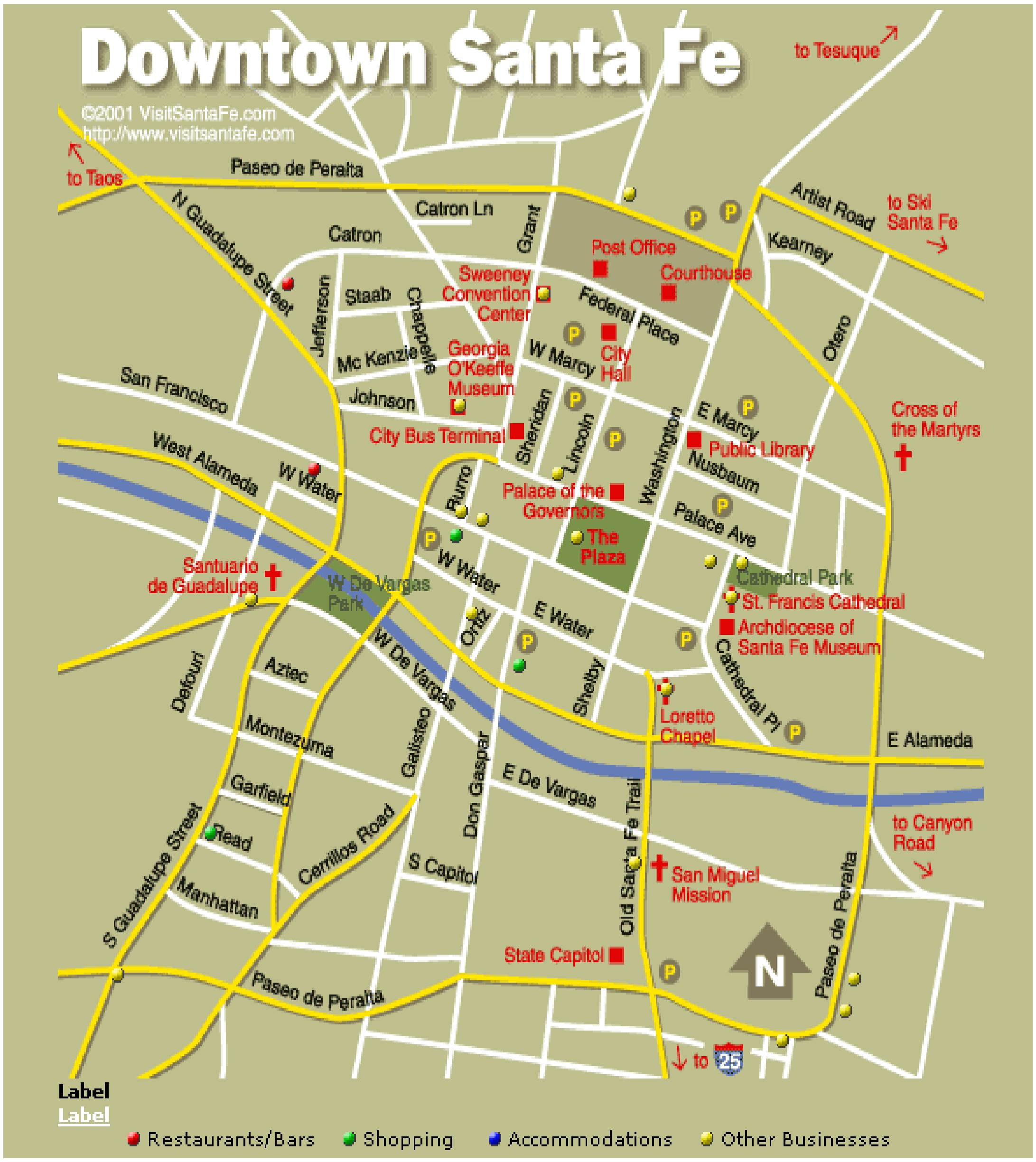
Alexander	Kozik	FF0040	akozik@atgc.org	University of California Davis Genome Center
Anelia	Kraltcheva	FF0041	anelia.kraltcheva@thermofisher.com	ThermoFisher/IonTorrent
Randall	Kramer		rkrm@novozymes.com	Novozymes
Ravi	Kumar		rvku@novozymes.com	Novozymes
Jochen	Kumm		jochen.kumm@gmail.com	IBM
Yuliya	Kunde		y.a.kunde@lanl.gov	Los Alamos National Laboratory
Patrick	Kwan		pkwan@cdc.gov	Centers for Disease Control and Prevention
Jennifer	Kwon		jkwon@lanl.gov	Los Alamos National Laboratory
Kurt	LaButti	FF0042	klabutti@lbl.gov	DOE Joint Genome Institute
Ernest	Lam	FF0043	elam@bionanogenomics.com	BioNano Genomics
Brad	Langhorst	FF0075	langhorst@neb.com	New England Biolabs
Ray	Langley		rlangley@lrri.org	Lovelace Respiratory Research Institute (LRRI)
Alla	Lapidus		piterlabs@gmail.com	St. Petersburg Academic University
Dominique	Lavenier	FF0044	lavenier@irisa.fr	IRISA, CNRS (Paris, France)
Huong	Le		huongl@amgen.com	Amgen
Tung	Le		tung0030@umn.edu	University of Minnesota
Darren	Lee		darren@nabsys.com	Nabsys Inc.
Darrin	Lemmer		dlemmer@tgen.org	Translational Genomics Research Institute, North
Eric	Lewis		elewis@tgen.org	Translational Genomics Research Institute, North
Po-E (Paul)	Li		po-e@lanl.gov	Los Alamos National Laboratory
Raphael	Lihana	FF0101, FF0102, FF0103	lihanaraphael@gmail.com	Kenya Medical Research Institute
Rebecca	Lindsey	FF0045	Rebecca.Lindsey@cdc.hhs.gov	Centers for Disease Control and Prevention
Chien-Chi	Lo	FF0046	chienchi@lanl.gov	Los Alamos National Laboratory
Chad	Locklear		clocklear@idtdna.com	Integrated DNA Technologies
Hrishikesh	Lokhande		hlokhande@ncgr.org	National Center for Genome Resources (NCGR)
Jon	Longmire		jonlongmire@lanl.gov	Los Alamos National Laboratory
Iain	MacCallum	FF0047	iainm@broadinstitute.org	Broad Institute
Mohammed-Amin	Madoui	FF0048	amadoui@genoscope.cns.fr	Genoscope/CEA
Punita	Manga	FF0049	pmanga@utk.edu	University of Tennessee, Knoxville / ORNL
Sophie	Mangenot		mangenot@genoscope.cns.fr	Genoscope/CEA
Michael	Marlowe		michael.marlowe@perkinelmer.com	PerkinElmer
Marta	Matvienko	FF0050	mmatvienko@clcbio.com	CLC bio - A Qiagen Company
Franklin	Mayanja		mayanjaf@gmail.com	Ugandan Ministry of Agriculture, Animal Industry and Fisheries
Carl	Mayers		cnmayers@dstl.gov.uk	Defence Science and Technology Laboratory (DSTL)
Kirsten	McCabe		kjmccab@lanl.gov	Los Alamos National Laboratory
Rebecca	McIntosh		rebeccam@lanl.gov	Los Alamos National Laboratory
Elizabeth	McLeod		lizm91@unm.edu	University of New Mexico
Benjamin	McMahon	FF0051	mcmahon@lanl.gov	Los Alamos National Laboratory
Timothy	McMahon		timothy.p.mcmahon10.ctr@mail.mil	Armed Forces DNA Identification Lab./American Registry of Pathology
Kimberly	McMurry		kmcmurry@lanl.gov	Los Alamos National Laboratory
Isaac	Meek		meek@neb.com	New England Biolabs
Robert	Mervis	FF0052	rmervis@gmail.com	CLC bio - A Qiagen Company
Ginger	Metcalf	FF0053	metcalf@bcm.edu	Baylor College of Medicine
David	Michaels	FF0054	dmichaels@clcbio.com	CLC bio - A Qiagen Company
Timothy	Minogue		timothy.minogue@us.army.mil	USAMRIID
Sam	Minot		sminot@signaturescience.com	Signatute Science, LLC
Karen	Moll		karen.moll@biofilm.montana.edu	Montana State University
Scott	Monsma	FF0055A, B	smonsma@lucigen.com	Lucigen Corp
Senzo	Mtshali		senzom@nicd.ac.za	National Institute for Communicable Diseases, South Africa
Joann	Mudge		jm@ncgr.org	National Center for Genome Resources (NCGR)
Mari	Murtskhvaladze		dna_lab@iliauni.edu.ge	Ilia State University
Joe	Musmacker	FF0056	joseph.musmacker@kapabiosystems.com	Kapa Biosystems
Donna	Muzny		donnam@bcm.edu	Baylor College of Medicine
Eugene	Myers	FF0057	myers@mpi-cbg.de	Max Planck Institute for Cell Biology & Genetics
Beth	Nelson		bane@novozymes.com	Novozymes
Judy	Ney		judy.ney@kapabiosystems.com	Kapa Biosystems
Peter	Ngam		pbn@ncgr.org	National Center for Genome Resources (NCGR)
Lisa	Nguyen		lisa@spiralgenetics.com	Spiral Genetics
Minh	Nguyen		Minh.Nguyen@usdoj.gov	National Institute of Justice
Christopher	Niblick		christopher.j.niblick.ctr@mail.mil	JPEO CBD
Jonathan	Nowacki		jnowacki@gmail.com	Roche Sequencing

Dawn	Obermoeller		dawno@biooscientific.com	Bioo Scientific
Juliette	Ohan		johan@lanl.gov	Los Alamos National Laboratory
John	Oliver	FF0058	oliver@nabsys.com	Nabsys Inc.
Christian	Olsen	FF0059, FF0060, FF0061	christian@biomatters.com	Biomatters' Geneious
Zbyszek	Otwinowski	FF0062	zbyszek@work.swmed.edu	University of Texas Southwestern Medical Center at Dallas
Oliver	Oviedo		oviedo@lanl.gov	Los Alamos National Laboratory
Suchitra	Pakala		sumisuchi@gmail.com	J. Craig Venter Institute
Suman	Pakala		suman.pakala@gmail.com	University of Georgia
Gustavo	Palacios		gustavo.f.palacios.ctr@mail.mil	USAMRIID
Beverly	Parson-Quintana		bapq@lanl.gov	Los Alamos National Laboratory
Lori	Peterson		peterson@cpsci.com	Caldera Pharmaceuticals (CPSCI)
Caleb	Phillips		caleb.phillips@ttu.edu	Texas Tech University
Thomas	Piggot		tjpiggot@dstl.gov.uk	Defence Science and Technology Laboratory (DSTL)
Marcella	Putman		Marcella.Putman@lifetech.com	Los Alamos National Laboratory
Thiru	Ramaraj	FF0063	tr@ncgr.org	National Center for Genome Resources (NCGR)
Teri	Rambo Mueller		teri.mueller@roche.com	Roche Sequencing
Brian	Raphael		BRaphael@cdc.gov	Centers for Disease Control and Prevention
Eric	Rees		eric.rees@researchandtesting.com	Research and Testing Laboratory
James	Robertson		james.robertson2@ic.fbi.gov	FBI Laboratory
Cooper	Roddey		cooper@edicogenome.com	Edico Genome, Corp.
Chandler	Roe	FF0105	croe@tgen.org	Translational Genomics Research Institute, North
Jeffrey	Rogers	FF0065	jr13@bcm.edu	Baylor College of Medicine
George	Rosenberg		ghrose@unm.edu	University of New Mexico
C. Nicole	Rosenzweig		carolyn.n.rosenzweig.civ@mail.mil	JPEO BMO
M.J.	Rosovitz		maryjo.rosovitz@nbacc.dhs.gov	National Biodefence Analysis and Countermeasures Center
Lori	Rowe		lrowe@cdc.gov	Centers for Disease Control and Prevention
Ashley	Sabol	FF0066	asabol@cdc.gov	Centers for Disease Control and Prevention
Jennifer	Saito		saitoj@hawaii.edu	University of Hawai
Antti	Sajantila		antti.sajantila@helsinki.fi	University of Helsinki, Finland
William	Salerno	FF0067	William.Salerno@bcm.edu	Baylor College of Medicine
Rashesh	Sanghvi	FF0068	rashesh.sanghvi@bcm.edu	Baylor College of Medicine
Faye	Schilkey		fds@ncgr.org	National Center for Genome Resources (NCGR)
Kelly	Schilling		kschilling@ncgr.org	National Center for Genome Resources (NCGR)
Matthew	Scholz		mscholz@msu.edu	Michigan State University - iCER
Andrew	Schuler		schuler88@gmail.com	University of New Mexico
Stephan	Schuster	FF0112	scs@bx.psu.edu	Nanyang Technological University Singapore
Anjali	Shah		Anjali.Shah@thermofisher.com	ThermoFisher/IonTorrent
Niranjan	Shekar	FF0069	niranjan@spiralgenetics.com	Spiral Genetics
CongCong	Shen		congcong@lanl.gov	Los Alamos National Laboratory
Xiaohong	Shen		xshen@lanl.gov	Los Alamos National Laboratory
Samuel	Shepard		vfn4@cdc.gov	Centers for Disease Control and Prevention
Palak	Sheth		psheth@bionanogenomics.com	BioNano Genomics
Heike	Sichtig	FF0114	Heike.Sichtig@fda.hhs.gov	US Food and Drug Administration (FDA)
Steve	Siembieda	FF0070	ssiembieda@aati-us.com	Advanced Analytical Technologies, Inc
Martin	Simonsen	FF0071	msimonsen@clcbio.com	Qiagen Aarhus
Gary	Simpson		garyl.simpson@comcast.com	University of New Mexico
Nick	Sisneros	FF0072	nsisneros@pacificbiosciences.com	Pacific Biosciences
Tom	Slezak		slezak@llnl.gov	Los Alamos National Laboratory
Jason	Smith		jrsmith@pacificbiosciences.com	Pacific Biosciences
Todd	Smith	FF0073	todd@digitalworldbiology.com	Digital World Biology
Shanmuga	Sozhamannan		shanmuga.sozhamannan.ctr@mail.mil	Critical Reagents Program (CRP)
Julie	Spencer		jaspence@unm.edu	University of New Mexico
Shawn	Starkenburg	FF0074	shawns@lanl.gov	Los Alamos National Laboratory
Fiona	Stewart	FF0098	stewart@neb.com	New England Biolabs
Doug	Storts		doug.storts@promega.com	Promega
Tod	Stuber	FF0076	Tod.P.Stuber@aphis.usda.gov	USDA-APHIS-NVSL
Anitha	Sundararajan	FF0077	asundara@ncgr.org	National Center for Genome Resources (NCGR)
James	Taylor		jtaylor2@dstl.gov.uk	Defence Science and Technology Laboratory (DSTL)
Ken	Taylor		ktaylor@aati-us.con	Advanced Analytical Technologies, Inc
Lee	Taylor		fflt@unm.edu	University of New Mexico
Robert	Taylor		rmtaylor@salud.unm.edu	University of New Mexico

Clotilde	Teiling		cteiling@illumina.com	Illumina, Inc.
Hazuki	Teshima		hazuki@lanl.gov	Los Alamos National Laboratory
Tea	Tevdoradze		t.tevdoradze@ncdc.ge	National Center for Disease Control and Public Health of Georgia
Sterling	Thomas	FF0099	Sterling.Thomas@noblis.org	Noblis
Graham	Threadgill		gjthreadgill@beckman.com	Beckman Coulter
Ruth	Timme	FF0097	ruth.timme@fda.hhs.gov	US Food and Drug Administration (FDA)
Masoud	Toloue	FF0079	mtoloue@biooscientific.com	Bioo Scientific
Chad	Tomlinson	FF0080	ctomlins@watson.wustl.edu	Washington University School of Medicine, The Genome Institute
Nicole	Touchet		peterson@cpsci.com	Caldera Pharmaceuticals (CPSCI)
David	Trees	FF0081	dlt1@cdc.gov	Centers for Disease Control and Prevention
Eija	Trees	FF0082	eih9@cdc.gov	Centers for Disease Control and Prevention
Angie	Trujillo		ATrujillo@cdc.gov	Centers for Disease Control and Prevention
Reem	Tubeishat		reem26484@yahoo.com	Jordan Univ. of Science and Technology, Princess Haya Biotechnology Ctr
Stephen	Turner	FF0083	sturner@pacificbiosciences.com	Pacific Biosciences
Eishita	Tyagi		vjn0@cdc.gov	Centers for Disease Control and Prevention
Joshua	Udall		jaudall@gmail.com	Brigham Young University
Susan	Ulanowicz		susan.ulanowicz@roche.com	Roche Sequencing
Pooja	Umale		peu@ncgr.org	National Center for Genome Resources (NCGR)
Sagar	Utturkar	FF0084	sutturka@utk.edu	University of Tennessee
Willy	Valdivia		willy.valdivia@orionbio.com	Orion Integrated Biosciences Inc.
Eric	Van Gieson		ericvangieson74@gmail.com	MRIGlobal
George	VanDegrift		gvandegrift@conveycomputer.com	Convey Computer
Narayanan	Veerarghavan	FF0086	narayanv@bcm.edu	Baylor College of Medicine
Eric	Vincent		eric.vincent@promega.com	Promega
Momo	Vuyisich		vuyisich@lanl.gov	Los Alamos National Laboratory
Darlene	Wagner		ydn3@cdc.gov	Centers for Disease Control and Prevention
Edward	Wakeland		edward.wakeland@utsouthwestern.edu	University of Texas Southwestern Medical Center at Dallas
Mark	Wang	FF0095	mwang@bcm.edu	Baylor College of Medicine
Kate	Weinbrecht	FF0087	kateldw@okstate.edu	Oklahoma State University
Richard	Wilson	FF0110	jpeck@genome.wustl.edu	Washington University School of Medicine, The Genome Institute
Richard	Winegar	FF0088	rwinegar@mriglobal.org	MRIGlobal
Aye	Wollam	FF0089	awollam@genome.wustl.edu	Washington University School of Medicine, The Genome Institute
Emily	Wong		emilyhwo@usc.edu	The University of Southern California
Matthew	Wong		matthew.wong@bcm.edu	Baylor College of Medicine
Jonathan	Wood	FF0090	jmdw@sanger.ac.uk	Wellcome Trust Sanger Institute
Gary	Xie	FF0091	xie@lanl.gov	Los Alamos National Laboratory
Lindsey	Yoder		lindsey.hollaway@okstate.edu	Oklahoma State University
Karina	Yusim	FF0092	kyusim@lanl.gov	Los Alamos National Laboratory
Ekaterine	Zhgenti	FF0093	eka_zh@hotmail.com	National Center for Disease Control and Public Health of Georgia
David	Zorikov		zorikov@gmail.com	National Center for Disease Control and Public Health of Georgia

NOTES

Map of Santa Fe, NM



History of Santa Fe, NM

Thirteen years before Plymouth Colony was settled by the Mayflower Pilgrims, Santa Fe, New Mexico, was established with a small cluster of European type dwellings. It would soon become the seat of power for the Spanish Empire north of the Rio Grande. Santa Fe is the oldest capital city in North America and the oldest European community west of the Mississippi.

While Santa Fe was inhabited on a very small scale in 1607, it was truly settled by the conquistador Don Pedro de Peralta in 1609-1610. Santa Fe is the site of both the oldest public building in America, the Palace of the Governors and the nation's oldest community celebration, the Santa Fe Fiesta, established in 1712 to commemorate the Spanish reconquest of New Mexico in the summer of 1692. Peralta and his men laid out the plan for Santa Fe at the base of the Sangre de Cristo Mountains on the site of the ancient Pueblo Indian ruin of Kaupoge, or "place of shell beads near the water."

The city has been the capital for the Spanish "Kingdom of New Mexico," the Mexican province of Nuevo Mejico, the American territory of New Mexico (which contained what is today Arizona and New Mexico) and since 1912 the state of New Mexico. Santa Fe, in fact, was the first foreign capital over taken by the United States, when in 1846 General Stephen Watts Kearny captured it during the Mexican-American War.

Santa Fe's history may be divided into six periods:

Preconquest and Founding
(circa 1050 to 1607)

Santa Fe's site was originally occupied by a number of Pueblo Indian villages with founding dates from between 1050 to 1150. Most archaeologists agree that these sites were abandoned 200 years before the Spanish arrived. There is little evidence of their remains in Santa Fe today.

The "Kingdom of New Mexico" was first claimed for the Spanish Crown by the conquistador Don Francisco Vasques de Coronado in 1540, 67 years before the founding of Santa Fe. Coronado and his men also discovered the Grand Canyon and the Great Plains on their New Mexico expedition.

Don Juan de Onate became the first Governor-General of New Mexico and established his capital in 1598 at San Juan Pueblo, 25 miles north of Santa Fe. When Onate retired, Don Pedro de Peralta was appointed Governor-General in 1609. One year later, he had moved the capital to present day Santa Fe.

Settlement Revolt & Reconquest
(1607 to 1692)

For a period of 70 years beginning the early 17th century, Spanish soldiers and officials, as well as Franciscan missionaries, sought to subjugate and convert the Pueblo Indians of the region. The indigenous population at the time was close to 100,000 people, who spoke nine basic languages and lived in an estimated 70 multi-storied adobe towns (pueblos), many of which exist today. In 1680, Pueblo Indians revolted against the estimated 2,500 Spanish colonists in New Mexico, killing 400 of them and driving the rest back into Mexico. The conquering Pueblos sacked Santa Fe and burned most of the buildings, except the Palace of the Governors. Pueblo Indians occupied Santa Fe until 1692, when Don Diego de Vargas reconquered the region and entered the capital city after a bloodless siege.

Established Spanish Empire
(1692 to 1821)

Santa Fe grew and prospered as a city. Spanish authorities and missionaries - under pressure from constant raids by nomadic Indians and often bloody wars with the Comanches, Apaches and Navajos-formed an alliance with Pueblo Indians and maintained a successful religious and civil policy of peaceful coexistence. The Spanish policy of closed empire also heavily influenced the lives of most Santa Feans during these years as trade was restricted to Americans, British and French.

The Mexican Period
(1821 to 1846)

When Mexico gained its independence from Spain, Santa Fe became the capital of the province of New Mexico. The Spanish policy of closed empire ended, and American trappers and traders moved into the region. William Becknell opened the l,000-mile-long Santa Fe Trail, leaving from Arrow Rock, Missouri, with 21 men and a pack train of goods. In those days, aggressive Yankeetraders used Santa Fe's Plaza as a stock corral. Americans found Santa Fe and New Mexico not as exotic as they'd thought. One traveler called the region the "Siberia of the Mexican Republic."

For a brief period in 1837, northern New Mexico farmers rebelled against Mexican rule, killed the provincial governor in what has been called the Chimayó Rebellion (named after a village north of Santa Fe) and occupied the capital. The insurrectionists were soon defeated, however, and three years later, Santa Fe was peaceful enough to see the first planting of cottonwood trees around the Plaza.

Territorial Period
(1846 to 1912)

On August 18, 1846, in the early period of the Mexican American War, an American army general, Stephen Watts Kearny, took Santa Fe and raised the American flag over the Plaza. Two years later, Mexico signed the Treaty of Guadalupe Hidalgo, ceding New Mexico and California to the United States.

In 1851, Jean B. Lamy, arrived in Santa Fe. Eighteen years later, he began construction of the Saint Francis Cathedral. Archbishop Lamy is the model for the leading character in Willa Cather's book, "Death Comes for the Archbishop."

For a few days in March 1863, the Confederate flag of General Henry Sibley flew over Santa Fe, until he was defeated by Union troops. With the arrival of the telegraph in 1868 and the coming of the Atchison, Topeka and the Santa Fe Railroad in 1880, Santa Fe and New Mexico underwent an economic revolution. Corruption in government, however, accompanied the growth, and President Rutherford B. Hayes appointed Lew Wallace as a territorial governor to "clean up New Mexico." Wallace did such a good job that Billy the Kid threatened to come up to Santa Fe and kill him. Thankfully, Billy failed and Wallace went on to finish his novel, "Ben Hur," while territorial Governor.

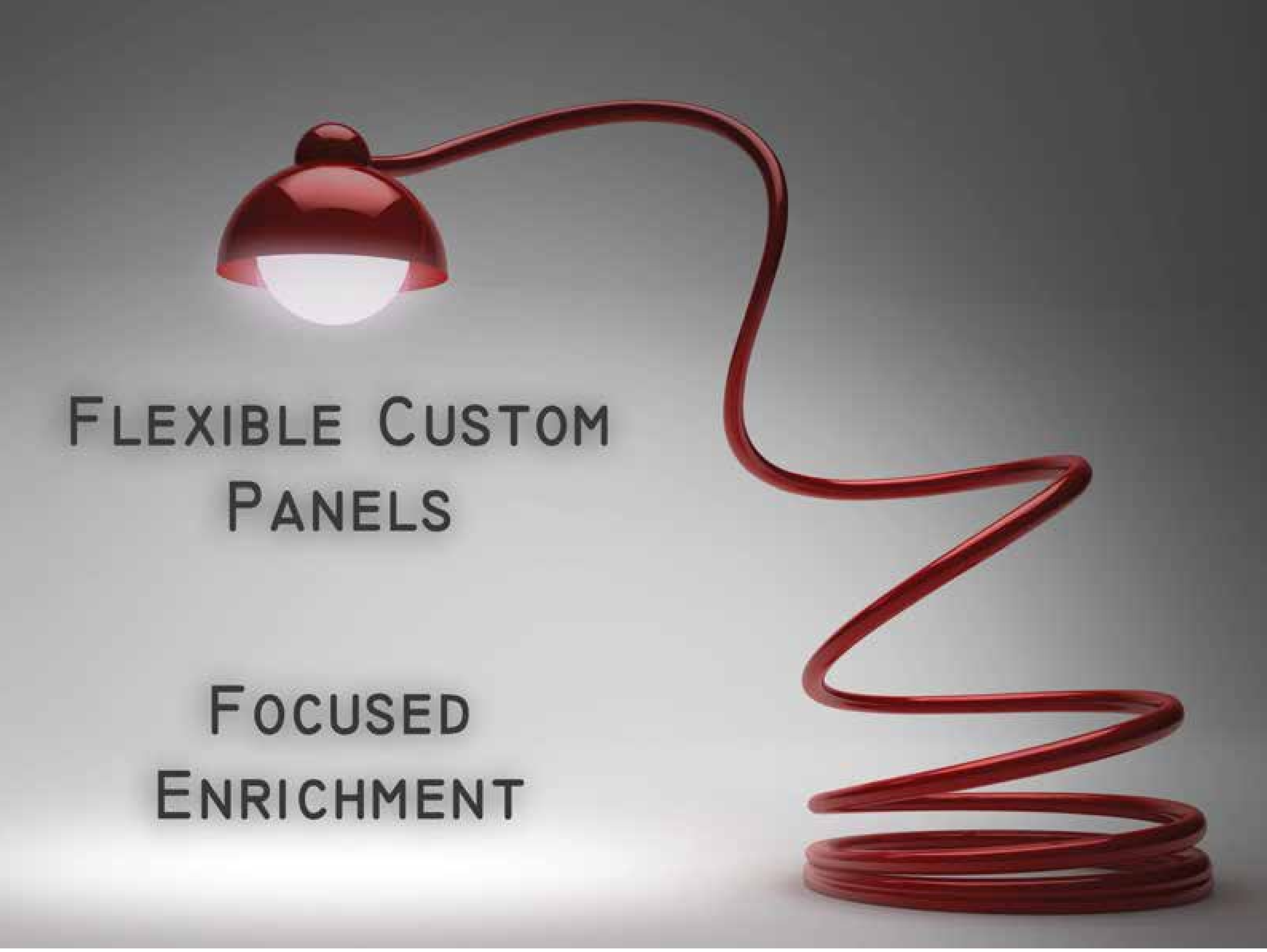
Statehood
(1912 to present)

When New Mexico gained statehood in 1912, many people were drawn to Santa Fe's dry climate as a cure for tuberculosis. The Museum of New Mexico had opened in 1909, and by 1917, its Museum of Fine Arts was built. The state museum's emphasis on local history and native culture did much to reinforce Santa Fe's image as an "exotic" city.

Throughout Santa Fe's long and varied history of conquest and frontier violence, the town has also been the region's seat of culture and civilization. Inhabitants have left a legacy of architecture and city planning that today makes Santa Fe the most significant historic city in the American West.

In 1926, the Old Santa Fe Association was established, in the words of its bylaws, "to preserve and maintain the ancient landmarks, historical structures and traditions of Old Santa Fe, to guide its growth and development in such a way as to sacrifice as little as possible of that unique charm born of age, tradition and environment, which are the priceless assets and heritage of Old Santa Fe."

Today, Santa Fe is recognized as one of the most intriguing urban environments in the nation, due largely to the city's preservation of historic buildings and a modern zoning code, passed in 1958, that mandates the city's distinctive Spanish-Pueblo style of architecture, based on the adobe (mud and straw) and wood construction of the past. Also preserved are the traditions of the city's rich cultural heritage which helps make Santa Fe one of the country's most diverse and fascinating places to visit.



Target Enrichment for Next Generation Sequencing

xGen® Lockdown® Probes for Target Capture

xGen® Lockdown® Probes are individually synthesized oligos for target enrichment, and custom panel development for sequencing. Each probe is assessed by mass spectrometry for quality control, enabling a seamless transition from discovery to clinical application.

- Complete and uniform enrichment with just 1X tiling
- Start projects faster with a 7–10 business day turnaround for 2000 probes
- Easily expand or optimize existing panels by adding custom probes
- Identify mutations, translocations, insertion sites, and more with a single enrichment technology

IDT®
INTEGRATED DNA TECHNOLOGIES

THE CUSTOM BIOLOGY COMPANY

WWW.IDTDNA.COM

100%

"Sponsors"

http://www.roche-diagnostics.us/
Meet and Greet Party

http://www.illumina.com/
Cowgirl Happy Hour

http://www.pacificbiosciences.com/
Lunch + lots extra!!!

http://www.promega.com
Lunch + extra!!!

“Sponsors”

NEW ENGLAND BioLabs Inc.

http://www.neb.com
Breakfast Each Day

IDT INTEGRATED DNA TECHNOLOGIES

http://www.idtdna.com/
Meeting Guides

Personalis®
Pioneering Genome-Guided Medicine

http://www.personalis.com
Keynote

Los Alamos NATIONAL LABORATORY EST. 1943

National Security Education Center

http://institute.lanl.gov/ias/
Poster Boards

MRIGlobal
National *Solutions* ***Worldwide*** *Impact*

http://www.mriglobal.org/
Lunch

“Sponsors”

sage science

http://www.sagescience.com
Keynote

ADVANCED ANALYTICAL

http://www.aati-us.com/
Keynote

DNAnexus

https://dnanexus.com/
Break

CLC bio
Accelerating Scientific Research

http://www.clcbio.com/
Break

BIONANO GENOMICS

http://www.bionanogenomics.com/
Break

KAPABIOSYSTEMS

http://www.kapabiosystems.com/
Break

ion torrent
by life technologies™

http://www.lifetechnologies.com
Lots of stuff!

PerkinElmer
For the Better

http://www.perkinelmer.com/
Lots of stuff!

BIOO SCIENTIFIC

http://www.biooscientific.com/
Lots of stuff!

"Sponsors"

Roche

illumina

Thank You !!!

PACIFIC BIOSCIENCES

MRIGlobal

National Solutions *Worldwide* Impact

sage science

Promega